UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Reference Tissue Normalization of Prostate MRI with automatic Multi-Organ Deep Learning Pelvis segmentation

Inês Correia Bagulho

Mestrado Integrado em Engenharia Biomédica e Biofísica
Engenharia Clínica e Instrumentação Médica

Dissertação orientada por:
Prof. Doutor Alexandre Andrade
Prof. Doutor Henkjan Huisman

2018

Acknowledgements

I would like to start by thanking the DIAG (Diagnostic Imaging Analysis Group) group at Radboud University Medical Center (Radboud UMC) for taking me in and making this experience memorable. I am thankful to be a part of such an important research group and to experience a real research environment. This project would not have been possible without the help of my external supervisor, Doctor Henkjan Huisman not only for providing me with this amazing opportunity but for his guidance and continuous support during this project. I am very grateful for all his advice, motivation and contagious enthusiasm. My thanks are also extended to my internal dissertation supervisor, Professor Alexandre Andrade. As my supervisor, his regular and prompt feedback, support and optimism were crucial in this work.
I would like to express my deepest appreciation to my fellow colleagues at the DIAG group who supported and helped me throughout this project, with a special thanks to Germonda Mooij who guided and advised me with her experience in research and programming. I would also like to express my gratitude to the Erasmus+ program for funding this internship.
Furthermore, I would like to thank all the amazing people I met during my internship abroad, in and outside of the office. Thank you for all the amazing experiences we lived together, from building a snowman to simply having lunch. With a special thanks to *Randomness in Nijmegen*, *DIAG - Lazy Lads*, and the roommates with whom I shared so much and had a great pleasure to live with. I am very thankful for the support of my very special friends, the ones who experienced the same 'suffering' along side me and the ones who cheered me on and motivated me to continue. A special thanks to my boyfriend Gonçalo Castilho who always maintained a smile on my face, even being more than 2000 km away.
Finally, this work was only made possible with the support of my parents who always showed me unconditional love and provided me with the best opportunities regardless of any obstacles facing their way. I also need to thank my two annoying older sisters without whom I would not be the person I am today. Without my family's help I would not have been able to, not only finish but even start this journey. I hope to be able to repay all the gifts I've been given and continue to share them with my family and friends.

Abstract

Prostate cancer is the most common cancer among male patients and second leading cause of death from cancer in men (excluding non-melanoma skin cancer). Magnetic Resonance Imaging (MRI) is currently becoming the modality of choice for clinical staging of localized prostate cancer. However, MRI lacks intensity quantification which hinders its diagnostic ability. The overall aim of this dissertation is to automate a novel normalization method that can potentially quantify general MR intensities, thus improving the diagnostic ability of MRI.

Two Prostate multi-parametric MRI cohorts, of 2012 and 2016, were used in this retrospective study. To improve the diagnostic ability of T2-Weighted MRI, a novel multi-reference tissue normalization method was tested and automated. This method consists of computing the average intensity of the reference-tissues and the corresponding normalized reference values to define a look-up-table through interpolation. Since the method requires delineation of multiple reference tissues, an MRI-specific Deep Learning model, Aniso-3DUNET, was trained on manual segmentations and tested to automate this segmentation step. The output of the Deep Learning model, that consisted of automatic segmentations, was validated and used in an automatic normalization approach. The effect of the manual and automatic normalization approaches on diagnostic accuracy of T2-weighted intensities was determined with Receiver Operating Characteristic (ROC) analyses. The Areas Under the Curve (AUC) were compared.

The automatic segmentation of multiple reference-tissues was validated with an average DICE score higher than 0.8 in the test phase. Thereafter, the method developed demonstrated that the normalized intensities lead to an improved diagnostic accuracy over raw intensities using the manual approach, with an AUC going from 0.54 (raw) to 0.68 (normalized), and automatic approach, with an AUC going from 0.68 to 0.73.

This study demonstrates that multi-reference tissue normalization improves quantification of T2-weighted images and diagnostic accuracy, possibly leading to a decrease in radiologist's interpretation variability. It is also possible to conclude that this novel T2-weighted MRI normalization method can be automatized, becoming clinically applicable.

Keywords: Magnetic Resonance; Normalization; Deep Learning; Reference-tissues

Resumo

O cancro da próstata é o tipo de cancro com maior incidência em homens que também constitui a segunda principal causa de morte por cancro em homens (excluindo o cancro de pele não melanoma). Uma área de intervenção tão importante como o cancro da próstata requer, inquestionavelmente, investigação constante que permita melhorar as técnicas de tratamento e diagnóstico já existentes ou possibilitar a criação de novas técnicas. A avaliação da gravidade das lesões em pacientes com cancro da próstata tem como propósito identificar e tratar somente os pacientes com doenças clinicamente significativas, comummente denominadas como doenças malignas.

A ressonância magnética da próstata é uma das modalidades de imagem médica que tem demonstrado crescente relevância na prática urológica, desde o início do seu uso clínico. Atualmente, a ressonância magnética está a tornar-se na modalidade de escolha para investigação do estágio de lesões prostáticas, permitindo uma boa diferenciação entre lesões clinicamente significativas (malignas) e lesões clinicamente não-significativas (benignas). No entanto, esta técnica de imagiologia ainda possui uma grande desvantagem, a variabilidade na escala de intensidades entre imagens, mesmo estas sendo adquiridas utilizando o mesmo scanner, sequência e paciente. Esta variabilidade de intensidades contribui para a variabilidade na interpretação de lesões, o que por sua vez dificulta a precisão do diagnóstico e posteriormente afeta também as medidas de tratamentos escolhidas.

Mesmo com o contínuo desenvolvimento da técnica de ressonância magnética da próstata, a identificação e classificação de lesões prostáticas focais utilizando ressonância magnética permanece um desafio devido a esta variabilidade de interpretação entre radiologistas, possivelmente causada pela variabilidade de intensidades. A normalização de imagens pode ser a solução, mais especificamente a utilização de técnicas de normalização que recorrem a tecidos de referência. Este tipo de normalização pressupõe que as oscilações de intensidade nos tecidos de referência, entre imagens, são causadas por dependências específicas dos scanners e parâmetros utilizados (denominadas como *machine dependencies* em inglês). Assim sendo, medir e corrigir essas dependências utilizando tecidos de referência deve permitir normalizar as imagens e quantificar as intensidades das mesmas. Desta maneira, as intensidades normal-

izadas devem ter uma precisão de diagnóstico aprimorada relativamente às intensidades originalmente adquiridas (*raw*).

Métodos que utilizam apenas um tecido de referência já demonstraram melhoramentos na precisão de diagnóstico, no entanto, é possível que estes métodos não consigam prever totalmente a distribuição não-linear das intensidades de ressonância magnética devido à utilização de uma única referência. Uma técnica de normalização que utilize múltiplas referências deve poder oferecer uma melhor aproximação desta distribuição de intensidades. Recentemente, foi demonstrada também uma melhoria na distinção entre lesões clinicamente significativas (PI-RADS 4-5) e lesões clinicamente não-significativas (PI-RADS 2-3) recorrendo a um método de normalização que utiliza múltiplos tecidos de referência. No entanto, esta abordagem proposta ainda apresenta uma forte limitação, visto que exige a segmentação manual de vários tecidos de referência, tornando o método inviável para uso clínico regular.

Deep Learning pode ser a ferramenta utilizada para automatizar a segmentação dos tecidos de referência, tornando este método aplicável regularmente em meio clínico. Técnicas de *Deep Learning* têm revolucionado muitos domínios de análise de imagens médicas, sendo um deles a segmentação de imagens médicas volumétricas. Por este motivo, uma variante do *state-of-the-art* de *Deep Learning* em segmentação de imagens volumétricas, denominado *U-net*, foi implementado neste estudo. O objetivo principal desta dissertação consiste em automatizar este novo método de normalização que pode potencialmente quantificar as intensidades gerais de ressonância magnética, melhorando assim a capacidade de diagnóstico da modalidade.

Duas bases de dados de ressonância magnética da próstata, adquiridas em 2012 e 2016 respetivamente, foram utilizadas neste estudo retrospetivo. Para melhorar a capacidade de diagnóstico de ressonância magnética ponderada em T2, um novo método de normalização que utiliza múltiplos tecidos de referência foi testado e automatizado. Este método consiste em calcular a intensidade média dos tecidos de referência e definir os valores de referência normalizados correspondentes para posteriormente definir uma curva de interpolação que relacione intensidades brutas (*raw*) e normalizadas. Como este método requer o delineamento de múltiplos tecidos de referência, um modelo de *Deep Learning* específico para ressonância magnética foi treinado em segmentações manuais e testado para automatizar essa etapa de segmentação. As segmentações automáticas, adquiridas com o modelo de *Deep Learning* mencionado, foram validadas e utilizadas numa experiência de normalização automática. O método de normalização com múltiplos tecidos de referência foi então aplicado, nas suas duas formas, manual e automática. Primeiro, a diminuição na variabilidade de intensidade entre as imagens após a normalização foi prevista e testada. Nesta experiência, a variabilidade de intensidade entre o conjunto de amostras de imagens brutas (*raw*) e normalizadas foi testada e comparada em todos os tecidos de referência, utilizando um teste F para igualdade de variância, com nível de significância de 0.05 (α = 0.05). Posteriormente, o efeito do método de normalização manual e automático na capacidade diagnóstica das intensidades ponderadas

em T2 foi determinado através de análises Característica de Operação do Receptor (ROC, *receiver operating characteristic* em inglês). As Áreas Debaixo da Curva (AUC, *area under the curve* em inglês) das duas abordagens (manual e automática) foram comparadas.

Experiências adicionais foram realizadas para testar a suposição inicial de que um método de normalização com múltiplas referências potencialmente forneceria uma melhor aproximação da distribuição não linear das intensidades de ressonância magnética comparativamente com um método de uma única referência. O método de normalização automatizado foi implementado utilizando entre 1 a 5 tecidos de referência (músculo Obturador Interno, Bexiga, Ossos Femur, Recto and Osso pélvico), aplicando todas as combinações de tecidos utilizados nesta dissertação e registando as AUC de cada implementação. As differentes AUCs, representativas das capacidades de diagnóstico das intensidades normalizadas, foram ilustradas num gráfico em função do número de referência utilizadas.

Em seguida todos os resultados obtidos foram apresentados e discutidos. A segmentação automática dos múltiplos tecidos de referência foi validada com um DICE *score* médio superior a 0.8 na fase de teste do modelo de *Deep Learning*. Posteriormente, o método de normalização desenvolvido foi implementado nas imagens manualmente e automaticamente segmentadas. Este método demonstrou uma diminuição de variabilidade significativa ($p_{value} < 0.05$) nos tecidos de referência após normalização, manual e automática, com exceção do tecido muscular do Obturador-Interno (denominado *Obturator-Internus muscle* em inglês). Subsequentemente, a diminuição na variabilidade de intensidades após normalização resultou numa melhor precisão de diagnóstico, usando tanto a abordagem manual, com uma AUC passando de 0.54 (*raw*) para 0.68 (normalizada), como a abordagem automática, com uma AUC passando de 0.68 (*raw*) a 0.73 (normalizada). As experiências adicionais demonstraram ainda que um aumento do número de referências resulta num aumento marginal na capacidade de diagnóstico das intensidades normalizadas. Este resultado não constitui um indício suficiente para corroborar a suposição inicial, no entanto, pode servir de indicação para futura investigação.

A normalização com recurso a múltiplos tecidos de referência demonstrou melhorar a quantificação das imagens ponderadas em T2 e a precisão do diagnóstico, possivelmente promovendo uma diminuição na variabilidade da interpretação dos radiologistas. Subsequentemente, os resultados apresentados nesta dissertação fornecem fortes provas de que as intensidades de imagens de ressonância magnética ponderadas em T2 estão correlacionadas com a malignidade da lesões e que métodos de normalização que recorrem a tecidos de referência têm um impacto positivo na precisão do diagnóstico do cancro da próstata. Tendo em conta os resultados adquiridos ao longo desta dissertação é ainda possível concluir que este novo método de normalização pode ser automatizado, tornando-se clinicamente aplicável, com potencial para ser aplicado a outras estruturas do corpo.

Palavras-chave: Ressonância Magnética; Normalização; Deep Learning; Tecidos de Referência

Contents

List of Figures

List of Tables

List of Abbreviations

AUC Area Under the Curve

CAD Computer-Aided Diagnosis

CNN Convolutional Neural Network

csPCa Clinically Significant Prostate Cancer

CZ Central Zone

DCE Dynamic Contrast Enhanced

DL Deep Learning

DRE Digital Rectal Examination

DWI Diffusion Weighted Imaging

LUT Look-Up Table

mp-MRI Multi-Parametric Magnetic Resonance Imaging

MR Magnetic Resonance

MRI Magnetic Resonance Imaging

PCa Prostate Cancer

PI-RADS Prostate Imaging Reporting and Data System

PSA Prostate-Specific Antigen

PZ Peripheral Zone

ReLu Rectified Linear Unit

ROC Receiver Operating Characteristic

SGD Stochastic Gradient Descent

T2W T2-Weighted Magnetic Resonance Imaging

TZ Transition Zone

Chapter 1

Introduction

1.1 Context

Prostate cancer is the most common cancer among male patients and second leading cause of death from cancer in men (excluding non-melanoma skin cancer) according to the study by Siegel et al. [1]. Therefore, it stands to reason that prostate cancer, as many other types of cancers, is one of the major societal challenges in health care. Since the causes of prostate cancer remain unknown, prostate cancer screening is the best solution to minimize its damage. The goal of prostate cancer screening is early detection and proper staging of prostate cancer which then allows a well-defined treatment. Magnetic Resonance Imaging (MRI) is currently becoming the modality of choice for clinical staging of localized prostate cancer [2]. However, MRI still has one major drawback, intensity variability which leads to the lack of a quantitative intensity scale, which in turn contributes to interpretation variability [3]. Even with recent developments in prostate MRI, identifying and classifying focal prostate lesions in MR imaging remains difficult, and the results obtained by experienced readers may not be achieved by less experienced readers [4].

Reference-tissue normalization [5, 6] could be a potential solution. It assumes inter-patient reference-tissue intensity variations to be caused by machine dependencies. Measuring and correcting these machine dependencies using reference-tissues should allow to normalize images. The normalized values should have an improved diagnostic accuracy over raw images. The one-reference tissue normalization method was shown to improve diagnostic accuracy [6], however, we believe a one-reference tissue model cannot fully predict the MR intensities non-linear distribution. A multiple-reference tissue method

should offer a better approximation. A pilot reference-tissue method was already shown to be successful in improving the discrimination of benign (PI-RADS 2-3) lesions from malignant (PI-RADS 4-5) lesions [7]. However, this approach requires a manual segmentation of multiple reference-tissues, making this approach not feasible for regular clinical use.

Recently, deep convolutional neural networks have revolutionized many medical image analysis domains, one of them being medical image segmentation with the U-net architecture [8] as the current state-of-the-art. Deep Learning could be the tool used to automate the reference-tissue segmentation, making it clinically applicable.

1.2 Dissertation Objectives and Outline

The overall aim of this dissertation is to automate a novel normalization method that can potentially improve the diagnostic ability of the current state-of-the-art in prostate cancer imaging, MRI, in a clinical setting. The multi-reference tissue normalization method proposed in this dissertation should allow automatic improvement of quantification of MR intensities, thus improving the diagnostic ability of MRI.

This dissertation is organized in 7 chapters described below. The present Chapter 1 introduces the context, motivation and general organization of the work.

Chapter 2 is subdivided in three sections. It starts with a brief explanation on prostate anatomy that allows an understanding over its multiple functional parts. The second section provides a comprehensive explanation of the relevant aspects surrounding the major topic that is prostate cancer. The third and final section introduces multi-parametric MRI as the current state-of-the-art in prostate cancer imaging, also describing its limitations and possible improvements.

Chapter 3 discusses the main intensity normalization strategies carried out in order to decrease, or even erase, the intensity variability present in MRI, with a special focus on reference-tissue normalization methods.

Chapter 4 provides a brief explanation on the Machine Learning branch denominated Deep Learning, with some detail on medical image segmentation and the current state-of-the-art architecture used for this task.

In Chapter 5, Materials and Methods used and developed in this dissertation are described. The multiple reference tissue method and Deep Learning architecture used are here explained in depth. Followed by the description of the dataset and an entire set of experiments that were carried out. The most relevant evaluation metrics or statistical methods used throughout this dissertation are also briefly explained here.

Chapter 6 presents the results obtained in this dissertation, after performing the experiments detailed in the Materials and Methods section, and the respective discussion.

In Chapter 7 a brief overview of the main findings and the overall conclusions of the entire dissertation are presented as well as future perspectives of this work.

Chapter 2

Prostate Gland

2.1 Prostate Anatomy

The prostate is a relatively small organ in the pelvis and a part of the male reproductive system, whose main function is to secrete prostate fluid, one of the components of semen. This walnut-sized gland is located between the pelvic bones, in front of the rectum and below the bladder. A schematic drawing of the pelvis and its different structures is shown in Figure 2.1.
A zonal classification of the prostate has been suggested by McNeal [10], as depicted in Figure 2.2, with three main functional parts: Peripheral Zone (PZ), Transition Zone (TZ) and (compressed) Central Zone (CZ).
The PZ is the largest area of the prostate and is mostly located at the back of the gland, closest to the rectal wall. The TZ is located centrally and surrounds the urethra, comprising approximately 5-10% of normal prostate volume. The CZ surrounds the ejaculatory ducts, comprising approximately 25% of normal prostate volume. The prostate can also be divided into 3 longitudinal portions, apex, median gland and base, where the apex refers to the lower part and the base to the upper part of the prostate.

2.2 Prostate Cancer

Prostate Cancer (PCa) is the most common cancer among male patients and second leading cause of death from cancer in men (excluding non-melanoma skin cancer) [1]. One in six men develop prostate

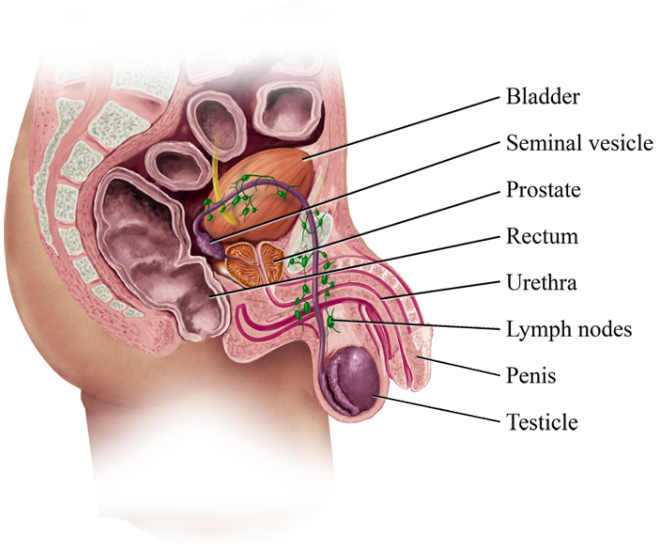


Figure 2.1: A saggital view drawing of the male pelvis in which several anatomical structures are indicated, with the Prostate located in front of the rectum and below the bladder. Source: [9]

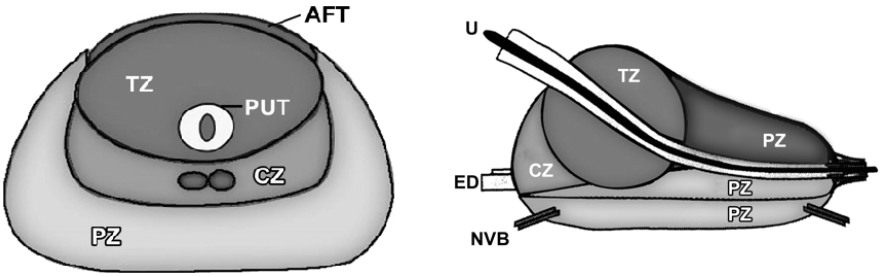


Figure 2.2: Prostate anatomy in transverse (left) and saggital (right) planes with division in different zones. AFT: anterior fibromuscular tissue, CZ: central zone, ED: ejaculatory duct, NVB: neurovascular bundle, PUT: periurethral tissue, PZ: peripheral zone, U: urethra, TZ: transition zone. Source: [11]

cancer at some point in their life [12]. Figure 2.3 shows the estimated number of new cancer cases and deaths in 2018 in the United States.

PCa growth can be characterized by two main types of evolution. The slow-growing or clinically non-significant tumours, comprising up to 85 % of all prostate cancers [13], usually keeping confined within the prostate gland. The treatment for these tumours can be substituted by active surveillance. The second variant of PCa, clinically significant PCa, progresses rapidly and the process of metastasis occurs from the prostate gland to other organs. Hence, trustworthy surveillance methods are crucial to differentiate between these two types of PCa. Initial detection of prostate cancer often occurs during routine physical exams. The common initial indicators of PCa are elevated Prostate-Specific Antigen (PSA) and an abnormal Digital Rectal Examination (DRE). PSA is a natural enzyme produced almost exclusively by the prostatic epithelial cells and is used as a biomarker for PCa detection. Although PSA screening

Estimated New Cases

Males				Females		
Prostate	164,690	19%		Breast	266,120	30%
Lung & bronchus	121,680	14%		Lung & bronchus	112,350	13%
Colon & rectum	75,610	9%		Colon & rectum	64,640	7%
Urinary bladder	62,380	7%		Uterine corpus	63,230	7%
Melanoma of the skin	55,150	6%		Thyroid	40,900	5%
Kidney & renal pelvis	42,680	5%		Melanoma of the skin	36,120	4%
Non-Hodgkin lymphoma	41,730	5%		Non-Hodgkin lymphoma	32,950	4%
Oral cavity & pharynx	37,160	4%		Pancreas	26,240	3%
Leukemia	35,030	4%		Leukemia	25,270	3%
Liver & intrahepatic bile duct	30,610	4%		Kidney & renal pelvis	22,660	3%
All Sites	**856,370**	**100%**		**All Sites**	**878,980**	**100%**

Estimated Deaths

Males				Females		
Lung & bronchus	83,550	26%		Lung & bronchus	70,500	25%
Prostate	29,430	9%		Breast	40,920	14%
Colon & rectum	27,390	8%		Colon & rectum	23,240	8%
Pancreas	23,020	7%		Pancreas	21,310	7%
Liver & intrahepatic bile duct	20,540	6%		Ovary	14,070	5%
Leukemia	14,270	4%		Uterine corpus	11,350	4%
Esophagus	12,850	4%		Leukemia	10,100	4%
Urinary bladder	12,520	4%		Liver & intrahepatic bile duct	9,660	3%
Non-Hodgkin lymphoma	11,510	4%		Non-Hodgkin lymphoma	8,400	3%
Kidney & renal pelvis	10,010	3%		Brain & other nervous system	7,340	3%
All Sites	**323,630**	**100%**		**All Sites**	**286,010**	**100%**

Figure 2.3: Estimates of new cancer cases and deaths (excluding non-melanoma skin cancer) in 2018 in the United States. Source: [1]

has been shown to improve early detection of PCa this biomarker is organ-specific, but not cancer-specific, as benign conditions may also cause elevated PSA values. In DRE, a physician palpates the prostate to search for irregularities. This technique is also not an accurate staging method due to difficult differentiation between benign and malignant tumors.

In case of a suspicious DRE and/or PSA level, a biopsy of the prostate is usually performed in order to establish the patient's definite diagnosis. The currently used technique for that is a transrectal ultrasonography–guided biopsy. The prostate tissue samples biopsied are then evaluated by the pathologist to determine whether PCa is present and, if so, the Gleason score of the cancer. The Gleason Score is the grading system used to determine the aggressiveness of prostate cancer and ranges from 2 to 10. Higher Gleason scores indicate the presence of a more agressive prostatic cancer. Gleason scores of 2-4 represent well differentiated or low-grade tumors. PCa with Gleason scores of 5-7 are labeled as moderately differentiated or intermediate grade. PCa with Gleason scores of 8-10 are labeled as poorly

differentiated or high grade. The current diagnostic approach, with PSA testing and DRE followed by transrectal ultrasonography–guided biopsy, lacks in both sensitivity and specificity in PCa detection and offers limited information about the aggressiveness and stage of the cancer [14]. Recent scientific work [2] shows the superiority of Magnetic Resonance Imaging, with or without targeted biopsy, over transrectal ultrasonography–guided biopsy, being the most sensitive and specific imaging tool for PCa diagnosis.

2.3 Prostate Magnetic Resonance Imaging

Prostate Magnetic Resonance Imaging is becoming the modality of choice for clinical staging of localized prostate cancer, finding 10% more clinically significant prostate cancers (csPCa) than outdated systematic biopsy [2]. The recommended technique of MRI for prostate cancer is multi-parametric Magnetic Resonance Imaging (mp-MRI). A typical prostate mp-MRI includes three-directional high-resolution anatomical T2-weighted MR imaging (T2W) in combination with functional MRI techniques such as diffusion-weighted imaging (DWI) and dynamic contrast enhanced (DCE) imaging.
Due to its high tissue contrast and spatial resolution, T2W can acquire information on not only prostate anatomy but also localization and staging of suspicious lesions. In T2W, the peripheral prostate zone is easily differentiated from the transition and central zone. The peripheral zone often appears with high signal intensity due to the high content of water in the glandular tissue opposed to the transition and central zone that often have lower signal intensity as shown in Figure 2.4.

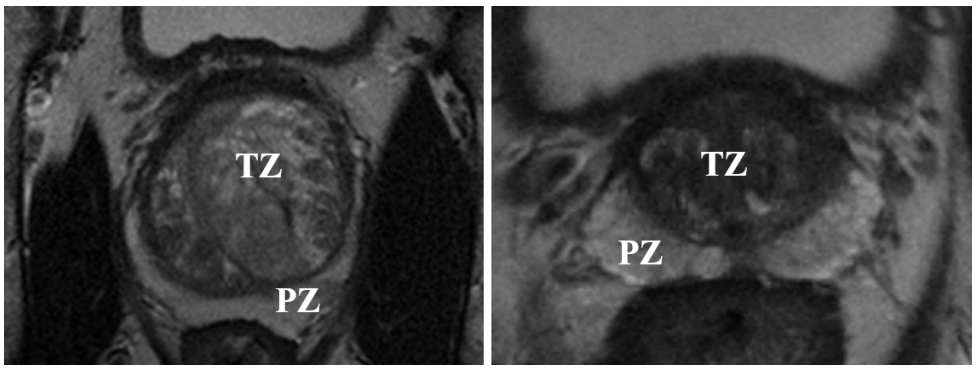


Figure 2.4: Axial T2-Weighted MRI slices of the male pelvis in which the Transition zone (TZ) and the Peripheral zone (PZ) are indicated.

PCa appears as an area of lower signal in T2W, since tumors fill up glandular space reducing water content compared to normal non-cancerous tissue. Low T2-weighted signals in the peripheral zone may also be seen in benign abnormalities, resulting in false-positive readings. Transition zone tumors are also very difficult to identify due to this zone heterogeneous appearance and presence of multiple benign

prostate hyperplasia nodules. Even so, T2W is considered to be superior of all the mp-MRI sequences for detection of cancer in the transition zone [15], and shows great potential in PCa staging and localization. A study [16] reported overall accuracies of 67%-69% for PCa localization using T2W imaging alone. By adding Diffusion weighted and dynamic contrast enhanced imaging to the T2-weighted images it was shown to significantly improve results on prostate cancer diagnosis [17]. Despite the recent developments in Prostate MRI, assessment of prostate lesions on mp-MRI is strongly dependent on the radiologist's expertise in interpreting MRI scans, and due to lack of standardization, also dependent on personal interpretation of image patterns and signal intensities. To tackle this issue, standardized guidelines were defined to provide strict steps to properly classify these lesions and improve reading reproducibility.

Prostate Imaging Reporting and Data System (PI-RADS) published by the European Society of Urogenital Radiology [18] and revised by the American College of Radiology [15] takes the location and size of a lesion into consideration and offers a decision process that results in a final five-point score to each lesion found in multiparametric MRI (T2W, Diffusion Weighted Imaging, Dynamic contrast enhanced), with 1 being a low-grade lesion and 5 being highly suspicious of malignancy.

The use of mp-MRI with this new lesion scoring system has increased the sensitivity in prostate cancer detection, however, the specificity remains poor, leading to over-treatment, such as unnecessary biopsies [19, 20]. It should be noted that biopsy is an invasive procedure which can result in serious infections or urine retention [21]. This poor specificity emphasizes the need for improvements to this technique to further improve the patient's chances of receiving a correct diagnosis and treatment. The improvements to this technique can be applied by focusing on improving its individual components, e.g. T2W. T2W is the basis of mp-MRI in prostate cancer diagnosis with its high tissue contrast and spatial resolution providing relevant information on not only prostate anatomy but also localization and staging of lesions. Even though T2-weighted MRI contributes significantly to prostate cancer staging and localization, it still has one major drawback, intensity variability that leads to the lack of a quantitative intensity scale, which in turn contributes to interpretation variability [3].

Chapter 3

Intensity Normalization

T2-weighted MR intensities are expressed in arbitrary units and do not have a tissue-specific value, meaning that the same tissue has a wide range of possible intensities, even when comparing images acquired with the same protocol, subject, and MR scanner. This lack of T2W intensity standardization influences the accuracy and precision of following image processing or medical analysis methods that rely on intensity similarity [22, 23]. MR intensity variability also contributes to inter-reader variability in a way that results obtained by experienced readers may not be achieved by less experienced readers [4].

Several intensity normalization algorithms have been proposed to bring MR intensities to a common scale and decrease this variability caused by machine dependencies.

3.1 Histogram-based Normalization

Nyul et al. [24] and Ge et al. [23] both proposed histogram-based normalization approaches. Currently, Nyul's method is the most frequently used normalization approach in MR imaging and is performed in two steps. The first step consists of a training stage to find the landmarks of the normalized gray scale. In this step the histogram is rescaled to the pre-set landmarks s1 and s2. These values are chosen by the user in a way that s1 and s2 should be larger than the expected minimum and maximum intensity of the subject histogram, respectively. The intensity values of p1 and p2 (0th and 98th percentiles) are determined and the rescaling is applied by mapping [p1, p2] to [s1, s2] linearly. Next, the foreground is determined by computing the overall mean intensity of the scan and applying that value as a threshold in the scan.

Then, the median of the foreground is taken as landmark. This procedure is repeated several times, for all histogram images, and the rounded mean of the determined median landmarks is taken as the standard median landmark (s50). The second step consists of a transformation stage to map the intensity values piece wise linearly with the obtained landmarks to the standard landmarks, thus acquiring the normalized images. It was later demonstrated [25] that lesion segmentation results were more accurate after applying this normalization method.

The reliable detection of the landmarks is a very challenging and essential task as the quality of the normalization heavily depends on it. Normally these landmarks are either chosen manually or are based on a segmentation, where pathology could have a great influence on the results. This is seen as the major drawback of this method.

3.2 Single Reference Tissue Normalization

To avoid the influence that pathology might have on the normalization method, studies investigated methods using a not-commonly cancerous reference tissue, such as the bladder lumen, levator-ani muscle or pubic bone. In one of those studies, Yahui Peng et al. [6] developed a reference tissue based normalization method using 4 different tissues to compare and define the optimal reference tissue for this implementation. The purpose of this study was to evaluate possible improvements in the classification of prostate cancer and normal prostatic tissue after applying the reference tissue intensity correction. The delineation of each tissue was needed and done manually. The normalized T2-weighted images were acquired using the ratio of the average raw T2-weighted signal by the average intensity value of the reference tissue.

A comparison between reference tissues and between scans from two scanners, Phillips and GE, were performed to investigate, respectively, the best tissue to be used as reference and to analyze possible discrepancies between the two vendors. The effectiveness of T2-weighted image signal intensities on differentiation between prostate cancer and normal prostatic tissue was estimated using the area under the Receiver Operating Characteristics (ROC) curve. Raw T2-weighted scans acquired from GE scanners reported better diagnostic ability compared to Phillips. However, the normalization method only improved the effectiveness of T2-weighted images from Phillips scanners. Phillips T2-weighted normalized images showed better results when compared with the non-normalized scans, regardless of the reference tissue implemented. The most consistent reference tissue used was the levator-ani muscle which showed the best results, especially in the Phillips T2-weighted images.

This method still has some limitations. It not only needs manual segmentation but also assumes a linear machine model, which is incorrect, since the influence of the MRI acquisition parameters on image intensities is nonlinear [26].

3.3 Sequence-based Normalization

In the work of Vos et al. [27] a method for normalizing T2-weighted images was developed. This method estimates the normalized T2-weighted values using not only the T2-weighted image and one reference tissue region, but also proton density values and a known sequence model, which are often not known or difficult to estimate accurately. The raw and normalized T2 weighted images are then input to a Computer-aided diagnosis (CAD) system. This CAD system analyzes each image, raw or normalized, and provides a diagnosis of the peripheral zone lesions found.

This Sequence-based Normalization method proposed by Vos et al., was shown to improve the CAD diagnosis of the system, in differentiating malignant peripheral zone from normal and benign peripheral zone. A significant improvement in discriminating performance was also achieved when differentiating normal peripheral zone from benign and malignant peripheral zone.

Despite the good results acquired with the proposed normalization approach, the requirements for the implementation of this method are too extensive to be applied to a big scale and in a general clinical setting, and cannot be applied retrospectively to standardly acquired T2WI.

3.4 Multi-reference tissue Normalization

Leung et al. [28] proposed a semi-automated segmentation technique that delineates cerebrospinal fluid, white matter and grey matter, tissues for which average intensities are computed. In a following step, linear regression between average intensities is performed and the results of this regression are used to acquire the normalized values. This study uses multiple reference tissues, obtaining more information regarding the MR intensity distribution compared to a one-reference tissue method. However, by performing linear regression it assumes a linear machine model which is incorrect. Scaling intensities with a simple linear transformation is not sufficient, as previously mentioned.

Recently, Stoilescu et al. [7] proposed a normalization method using multiple reference tissues which does not require detailed knowledge about the sequence model, nor the need for an additional sequence, achieving similar results to the Sequence-based normalization method study previously mentioned. This approach does not assume a linear machine model. The proposed method consists of fitting a smooth spline model to the (tissue average intensity, tissue T2 relaxation time) pairs. The map of the normalized values is acquired inverting the resulting spline. The normalized T2 weighted intensities are then input to a CAD system responsible for the discrimination of malignant (PI-RADS 4-5) and benign (PI-RADS 2-3) lesions. This novel normalization method was shown to improve significantly the diagnostic accuracy of CAD system.

With this study it was possible to normalize intensities taking into consideration their non-linear distribution across the MR scan. However, before implementing this method, a manual segmentation of four reference tissues in the Pelvic Region (gluteus maximus, body fat, femoral head, bladder lumen) was needed, which makes this an approach not feasible for regular clinical use. Deep Learning could be the tool used to automate this multi-reference tissue segmentation.

Chapter 4

Deep Learning

Similarly to humans, in order to comprehend images, computers require a set of discriminative information, commonly called features. Machine learning algorithms generally learn to map a given feature vector to the output labels. However, designing handcrafted methods that extract all the optimal features is a commonly arduous task and results in moderate performance. Instead of using this additional methods or handpicking the features, a more efficient method is to allow the algorithm to extract the optimal set of features from the data. Deep learning is a branch of machine learning, which allows computer systems to learn features directly from the input data. This fast-growing field uses statistical techniques to create autonomous and self-teaching systems. A Deep learning model consists of a set of feature extraction layers and classification layers [29]. Each layer learns increasingly abstract and complex features from the raw input, such as pixels from images. In the first layer simple features, such as edges, are identified based on the input. Thereafter, in the second layer more complex features are identifies, like corners and contours, given the output of the previous layer. Going deeper into the network more complex and abstract features are encoded. The final layer returns a classification output which indicates the probability of the input data belonging to a certain class.

4.1 Artificial Neural Networks

Artificial neural networks (or just neural networks) were inspired by the biological neural system. Neural networks are modeled as a structure of neurons connected in layers. These architectures usually consist of three different layers: the input layer that contains the input feature vector; the output layer which

consists of the neural network final response; and the layer in between that contains the neurons (nodes) that connect to both the input and output, as shown in Figure 4.1.

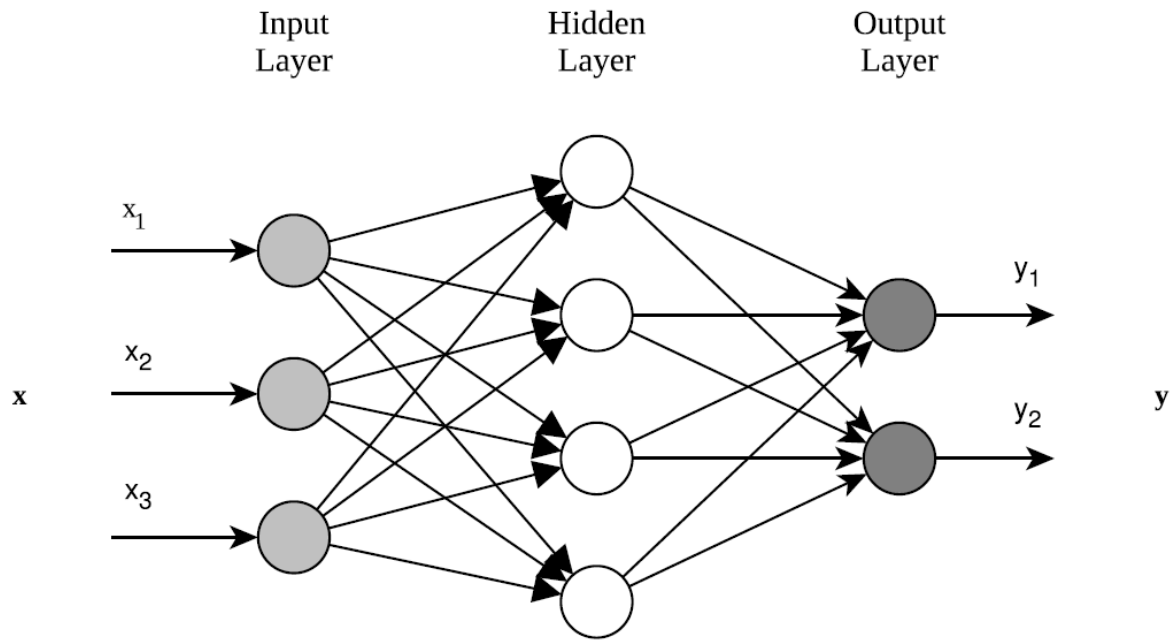


Figure 4.1: An example of a neural network with three layers. The input layer receives the input data, a vector x of N elements commonly called neurons, in this case N = 3. The hidden layer then computes a new representation of the input. Finally, the output layer computes the answer y of the network from the hidden representation.

It is important to note, that there can be several hidden layers in a network, with the term "deep" learning implying multiple hidden layers. The neural network represented in Figure 4.1 only allows signals to travel from input to output, making it a feed-forward neural network. A neural network is typically characterized by three main features: the network architecture; input and activation functions; and the weight of input connections. The architecture of the network is defined by its number of layers, its number of neurons in each layer and the connections established between the neurons. The neuron is the basic unit of computation in a neural network. Each neuron has an associated value which is either given by the data, for the input layer, or computed from the set of its inputs, provided by the previous layer. Before being passed to the following layer, the output of a layer is usually first input to an activation function that defines the mathematical function that is used to compute the state of a neuron based on the values of its inputs. The main purpose of the activation function is to introduce non-linear properties to the network that allow the network to not just learn and compute a linear function but to have a higher degree of complexity learned. The output of a layer is then connected, through synapses, to neurons from the next layer, as seen in Figure 4.1. These connections between neurons have an associated weight, which can be thought of as the strength of the input in determining the output. The forward pass, which computes values from inputs to output, is finished and a backward pass begins. The backward pass refers to the process of learning, using an optimization algorithm. The learning procedure

or optimization algorithm, the final characteristic of a neural network, is applied to optimize the weights of the network, in order to approximate the desired function as well as possible. The algorithm gives a prescription for adjusting the weights. The Stochastic Gradient Descent (SGD) algorithm is one of many optimization methods based on the analysis of the gradient of the objective. This training algorithm consists of a five-step process. Initially, the input is given to the network. Thereafter, the system output is compared to the expected result and an error is computed for each neuron. A specific loss function is chosen to minimize the difference between the input and desired output. Then, the gradients of the computed errors are applied to the weights of the neurons and these weights are updated so that the chosen loss function value decreases. This process is then repeated for each example of the data set and then iterated again as long as the intended loss classification error is not reached. One single application of the entire data set is called an epoch. The optimal weight arrangement is obtained through multiple adjustments (iterations) of the weights by the algorithm. The learning rate is the parameter that controls how much the weights of the network are adjusted with respect to the loss gradient. The lower the learning rate the slower the training convergence. The progression of the learning of a neural network is referred to as convergence. The Adam optimization algorithm, nomenclature derived from adaptive moment estimation, is an extension to SGD that has recently been used in deep learning applications. The Adam optimizer is very similar to the Stochastic gradient descent, however, instead of maintaining a single learning rate for all weight updates it computes and uses individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients

It is also important to note that, before starting the learning process of the network, a proper initialization of the weights is required to speed up the convergence process.

The architecture of the network plus its trained weights is considered as the fully trained model. The data used in the training phase is called training dataset, while the data that will be used to evaluate the performance of the estimated model is called test set. The test phase, where the test set is applied, consists of comparing the predicted output with the truth and evaluating the predictive ability of the trained model. Before the test phase, usually some parameters of the network need to be tuned for the network to reach optimal results. An additional dataset, called validation dataset, is used for this purpose since the test set should only be used in the final performance analysis and remain untouched until then. Using the test set to tune the parameters of the network could cause the model to be custom-made for those set of samples and not be generalized for other independent samples, problem denominated overfitting.

Overfitting occurs when the model has learned the data in such a great extent that it fails to capture underlying general information. The network must learn a general solution instead of a solution that is too tightly coupled with the training examples, a poor generalization causes this problem known as overfitting of the training data. Since the purpose of these deep learning models is not to classify the training samples, but rather to classify samples from an independent test set composed of unseen data, it is necessary that the model learns general information that can be applied to both the training and the test dataset. The size of the training data can have a large impact on this overfitting factor.

To avoid overfitting of the training data, several regularization techniques are used. One way to approach this problem is to define an appropriate (not too large) number of training epochs, constraining the weights to not grow too large. The generation of artificial samples from the existing training samples, approach called data-augmentation, is another very efficient technique that can solve overfitting [30], increasing the amount of input data by applying minor changes to the existing dataset. L_2 regularization is also one of the most common regularization techniques, adding a contribution to the loss function to avoid overfitting.

4.2 Convolutional Neural Networks

Convolutional neural networks (CNN) are a sub-class of deep learning and a type of artificial neural networks whose main calculation operation is convolving filters over matrices. It is important to note that neurons in Artificial neural networks commonly correspond to filters or kernels in CNNs. In the simplest case, a CNN architecture is a list of layers that transform the image volume into an output volume. Traditional CNNs have three important building tools: Convolutional layers; Pooling layers; Fully Connected Layer.

4.2.1 Convolutional Layer

The Convolutional layer is the core building block of a Convolutional neural network. The Convolutional layer's parameters consist of a set of learnable filters. Each of these filters can be thought of as feature identifiers. The input data specifies the width, height, and number of channels. Every filter is small spatially (compared to the input) along width and height, but extends through the full depth of the input volume. If the number of channels of the input is three, then the filters will have the same number of channels. Typically, the number of channels is commonly three for images, for the RGB values for each pixel.

Each filter is convolved across the width and height of the input volume and dot products are computed between the region of the filters in the input layer and the weights to which they are locally connected in the output layer, as shown in Figure 4.2.

The sliding of the filter over the input allows the computation of the entire output volume. A parameter, called stride, is used to control the sliding of the filters. For instance, if the stride is equal to 1, then the filters are moved one pixel at a time in the convolutional layer. And if the stride is equal to 2, then the filters jump 2 pixels at a time as they are being slid through the input. Another dimension would be added to the stride in order to apply it to convolution example in Figure 4.2. The increase of the number of filters used, would consequently increase the number of feature maps learned. However, the increase of the stride length would allow the reduction of the output volumes spatially (along width and height).

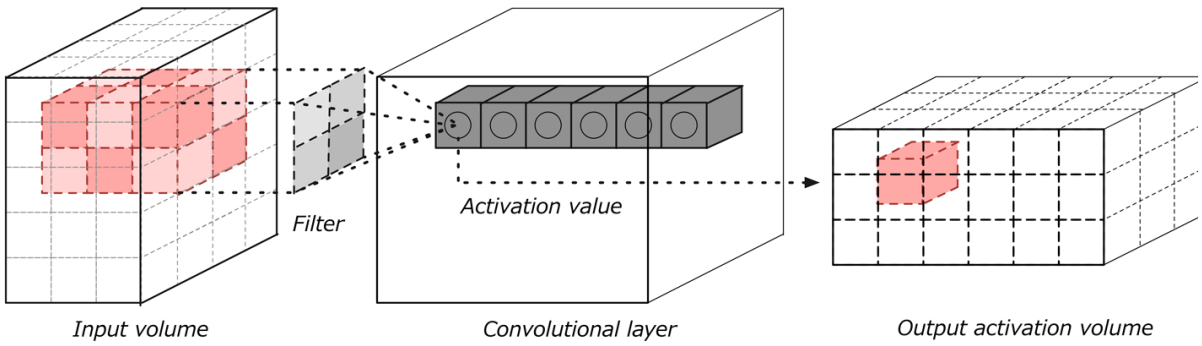


Figure 4.2: Representation of a convolution layer with input and output volumes. At each of the filter's location a dot product between the values of the filters (weights) and the coinciding input arrays are computed. The calculated output value is then assigned to the respective location in the output volume. Source: [31]

In the backward stage of a convolutional layer, the updating of the weights of the filters is also done through back-propagation. The back-propagation process for a convolution operation is also a convolution, but with spatially-flipped filters.

4.2.2 Pooling Layer

Convolutional layers are often followed by pooling layers that reduce the spatial size of the representation to reduce the amount of parameters and computation of the network. Pooling layers are, thus, also known as downsampling layers. No learning takes place in the pooling system. Max-pooling and Average pooling are some of the most common types of pooling layers currently used. Pooling layers can be compared to Convolutional layers. In pooling layers, instead of producing the dot product with the underlying matrix, the area at hand is taken and one value is computed (the operation used depends on the type of layer). In Max-pooling and Average pooling, the maximum and the average are respectively computed. Thus, the parameters, filter and stride lengths, used in the pooling layer define the amount of downsampling done. These pooling layers also play an important role in allowing the decrease of chance of overfitting.

4.2.3 Fully Connected Layer

A Fully Connected Layer, or dense network, has the most classical topology, having full connections between its neurons and all activations in the previous layer. In classification models the final layer is usually a fully connected layer and there are as many output nodes as there are possible classes. This means that every activation map value from previous layer has a weighted connection to each output layer node. In segmentation tasks, the last layer of the model is usually convolutional.

4.3 U-net

Recently, deep learning has revolutionized many medical image analysis domains, one of them being medical image segmentation with the U-net architecture [8] as the current state-of-the-art. Çiçek et al. [32] picked up Ronneberger et al. [8] work and applied the U-net idea on a 3D implementation, creating the 3D U-net.

The network does not include the usual stacking of layers. The 3D U-net, like the standard U-net [8] consists of symmetric analysis (descending) and a synthesis (ascending) path each with four resolution steps. The analysis part reduces spatial dimensions of features maps. The synthesis part recovers details and spatial dimensions. In the analysis path, each layer contains two 3x3x3 convolutions each followed by a rectified linear unit (ReLu) activation funtion, and then a 2x2x2 max-pooling with strides of two in each dimension. In the synthesis path, each layer consists of an up-convolution of 2x2x2 by strides of two in each dimension, followed by two 3x3x3 convolutions each followed by a ReLu, as shown in Figure 4.3. The ReLu is one of the most commonly used activation functions in deep learning architectures and is defined by f(x) = Max(0, x), x being the input of the neurons.

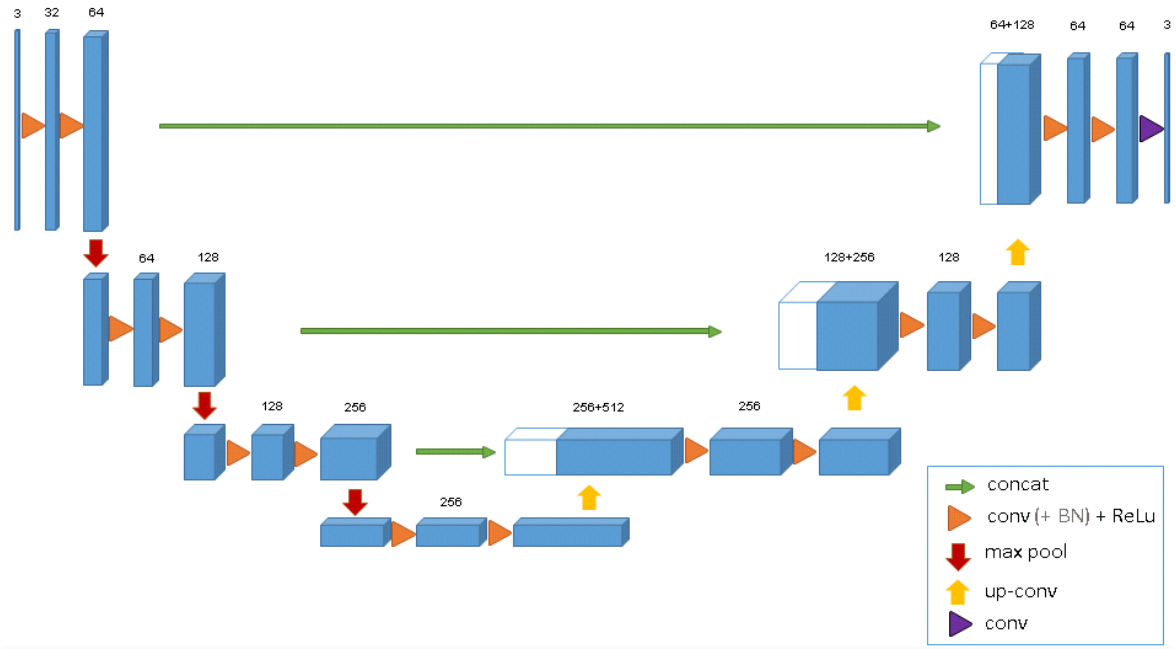


Figure 4.3: The 3D u-net architecture. The blue boxes represent feature maps with the number of channels denoted above each feature map. The white boxes correspond to feature maps that were calculated in the analysis path and concatenated with the synthesis path. The green arrows represent skip connections. The up-convolutional layers represent a special convolution operation that increases the activation map size. Source: [32]

The so-called skip connections, shown as green arrows in Figure 4.3, combine every two layers of the same resolution to avoid the loss of image information. This bridge between the analysis and synthesis

paths was implemented to allow the synthesis path to gain the essential high-resolution features from the analysis path, thus improving the output contours of the network. In addition, these connections also speed up the convergence of the network.
This network architecture is designed so that it is able to obtain accurate segmentation using only a few training images.

4.4 Anisotropic 3D U-net

Recently, Mooij et al. [33] created a novel MRI-specific multi-organ segmentation method 3D U-net, by adjusting the Çiçek's [32] 3D U-net architecture architecture to reflect the anisotropy in the dimensions of the MRI image volumes. MRI scans are image volumes with high resolution in one plane (axial plane), while in the third direction the slices are much thicker. Hence, these anisotropic image volumes can be thought of as just stacks of 2D images. This multi-organ segmentation architecture is very similar to the original 3D U-net, with the difference that it starts and ends with two layers of two 2D convolutions and one 2D max-pooling each, to consider the anisotropy of the input MRI volumes (stack of 2D images), as shown in Figure 4.4.

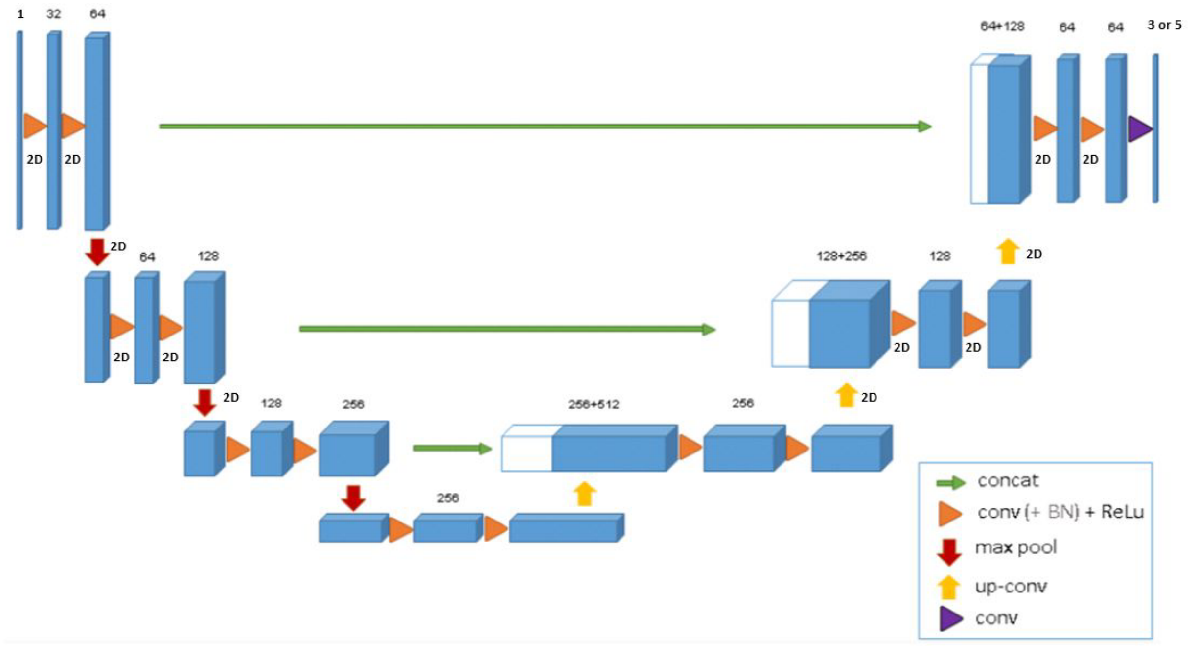


Figure 4.4: MRI-specific Deep Learning architecture, called aniso-3DUNET developed by Mooij et al. [33] and originally based on Çiçek's [32] 3D U-net architecture.

The final step to this network is a softmax either to 3 or 5 labels. The softmax function is often used in the final layer of a neural network-based classifier. Softmax is used in multi-label classification tasks,

assigning decimal probabilities, that add up to 1, to each class. This architecture was applied to segment prostate zones (peripheral and transition) and showed marginal improvements when compared to a recently published ATLAS method [34].

Chapter 5

Materials and Methods

5.1 Multi-Reference Tissue Normalization Method

The Multi-Reference Tissue Normalization method assumes that the MRI signal intensities are mainly dependent on tissue properties and machine dependencies. It also assumes that reference-tissues are not affected by disease and have MRI-specific tissue properties that are similar between patients. The reference-tissues can then be used to characterize the machine dependencies. First, the segmentation, either manual or automatic, of the reference-tissues is needed, so that the average intensity of each reference-tissue can be collected and assigned a normalized pre-defined value that is tissue-specific and generalized for all patients. Before computing the averages of the tissues, a pre-processing step of 3-pixel erosion was applied to the segmented reference-tissues to ensure a well-defined region of interest. Having computed all (reference-tissue average intensity, normalized reference value) data-points, a Look-Up Table (LUT) is defined through interpolation of the reference-pairs. Then, the raw image intensity is input to the LUT and the normalized map is computed. In short, the entire method is shown in Figure 5.1.

The continuous normalization LUT was constructed using interpolation between the discrete data-points. The Multi-Reference Tissue Normalization method assumes that the mapping function is monotonously increasing. However, the assumption that reference tissues gray values always appear in the same order relatively to each-other (e.g. muscle with the lowest gray value, followed by bone and so forth) may not always be correct, which can cause the (reference-tissue average intensity, normalized reference value) data-points to not be monotonously increasing. The data is first optimized with an isotonic regression

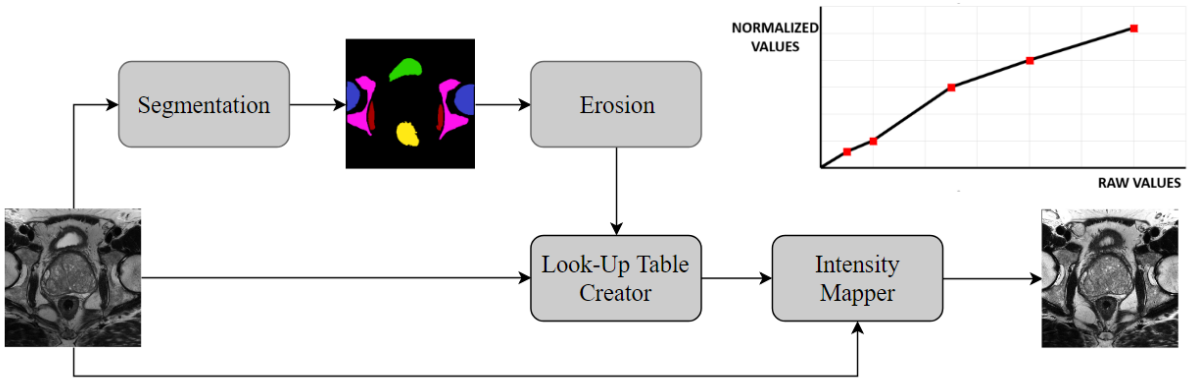


Figure 5.1: Diagram explaining the Multi-Reference Tissue Method applied in this study. This process starts with a T2-weighted raw and finishes with the corresponding normalized T2-weighted image.

method, that uses the standard-deviation inverse of the data points as weights to enforce monotonicity. Then, the continuous normalization LUT was constructed using interpolation between the output monotonously increasing discrete data-points. Thus, a piece-wise monotonously increasing model was fitted. Two piece-wise interpolation methods were used, Linear and Cubic Hermite interpolation. This method was fully implemented by the author of this dissertation using python as the programming language.

5.1.1 Linear Interpolation

Linear interpolation is the most basic interpolating function consisting of connecting straight lines between neighbouring data points. For a set of data points (x_k, y_k), with k = 1, 2 , ., ., n piece-wise linear interpolation satisfies:

$$P(x_k) = y_k \qquad P(x_{k+1}) = y_{k+1} \tag{5.1}$$

The linear interpolant at any point x, with $x_k < x < x_{k+1}$, takes the form:

$$P(x) = y_k + (x - x_k)\frac{(y_{k+1} - y_k)}{(x_{k+1} - x_k)} \tag{5.2}$$

5.1.2 Cubic Hermite Interpolation

Cubic Hermite interpolation method is typically used for interpolation of numeric data, requiring both the values and the derivatives of an interpolating function to fit the given data. For a set of data points (x_k, y_k), with k = 1, 2 , ., ., n piece-wise cubic Hermite interpolation satisfies:

$$P(x_k) = y_k \qquad P(x_{k+1}) = y_{k+1} \tag{5.3}$$

$$P'(x_k) = y'_k \qquad P'(x_{k+1}) = y'_{k+1} \tag{5.4}$$

Whereas with piecewise linear interpolation two conditions led to a linear interpolant, four interpolating conditions, mentioned above, give a cubic interpolant. Using piece-wise Cubic Hermite interpolation, two neighbouring data points should be taken into account at a time, x_k and x_{k+1}. The cubic Hermite interpolant at any point x, with $x_k < x < x_{k+1}$, takes the form:

$$P(x) = \frac{3hs^2 - 2s^3}{h^3} y_{k+1} + \frac{h^3 - 3hs^2 + 2s^3}{h^3} y_k + \frac{s^2(s-h)}{h^2} d_{k+1} + \frac{s(s-h)^2}{h^3} d_k \tag{5.5}$$

where

$$h = x_k + 1 - x_k \tag{5.6}$$

$$d_k = P'(x_k) \tag{5.7}$$

$$s = x - x_k \tag{5.8}$$

Piece-wise cubic Hermite interpolation can prove very useful in fitting a smooth, continuous and monotonous model.

5.1.3 Reference Tissues

Reference-based normalization methods can have either one or multiple references. Even though using a one-reference tissue approach reduces complexity of the method, we believe it still cannot fully predict the MR intensities non-linear distribution. The non-linear MR intensity distribution can, however, be approximated by a multi-point curve. Thus, the proposed multi-reference tissue normalization method could be the solution. The choice of reliable reference tissues is an essential task as the quality of the normalization strongly depends on it. The optimal reference-tissues should provide enough information over the entire gray value distribution while satisfying homogeneity requirements. The reference-tissues should also not be affected by disease. For this study in particular, closeness of the tissue to prostate intensity values should also be taken into account in order to increase the knowledge over prostate-tissue intensities.

The aforementioned tissue characteristics were considered in the choice of reference tissues used in this study. Tissues from both ends of the signal intensity distribution of the T2-weighted images were chosen, the bladder lumen with a high signal intensity and the obturator-internus muscle with low signal intensity, to try to cover the whole spectrum of intensities in the scans. Five reference tissues were used in this study: Bladder lumen, the Femur heads, the Pelvic bone, the Rectum and the Obturator-Internus muscle, identified in Figure 5.2.

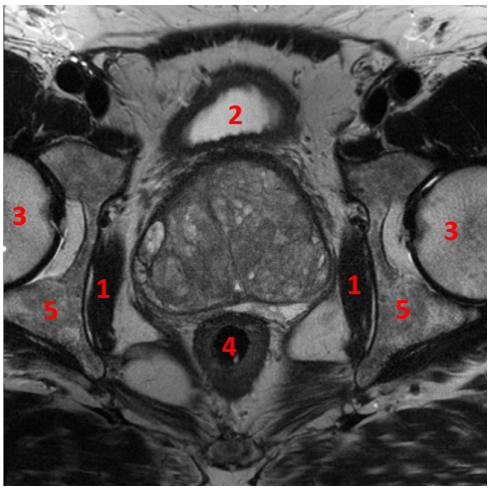


Figure 5.2: T2-weighted prostate MRI axial slice with the five reference-tissues identified: 1- Obturator-Internus muscle, 2- Bladder lumen, 3- Femur heads, 4- Rectum, 5- Pelvic bone

5.2 Deep Learning - Multi-Reference Tissue Segmentation

Deep Learning has been shown to be successful in several medical image segmentation tasks. An MRI-specific multi-organ segmentation method proposed by Mooij et al. [33] was implemented in this study to automate the segmentation of reference-tissues. As mentioned in Chapter 4, this novel 3D U-net variant, called aniso-3DUNET, reflects the anisotropy of the input MRI volumes (stack of 2D images) in its architecture.

Since the purpose is to classify the five reference-tissues, the final step in the implementation of this architecture is a softmax to 6 labels (background, Obturator-Internus muscle, Bladder lumen, Femur heads, Rectum, Pelvic bone), preceded by two steps for each voxel that map 64 by 64, 64 by 6 features, as shown in Figure 5.3.

The original image resolution of 0.5x0.5x3.6 mm and 0.3x0.3x3.6 mm was resampled to 1x1x3.6mm, in order to fit the full images into GPU memory, since the runs were conducted in an NVIDIA GPU. The deep learning model was trained on a set of 32 T2-weighted transversal volumes and tested on a set of 8 T2-weighted transversal volumes. The training and testing data were completely separated. During the training phase, data-augmentation was applied through small 3D rotations and left-right flips, in order to avoid overfitting and to increase the input training data. Due to the limited data available for training, validation scores were expected to vary depending on the validation split, so a 4-fold cross validation was carried out to obtain a more reliable range of validation scores. The models were trained with a learning rate of 0.0001, a standard type of normalization in medical imaging, called glorot uniform initialization, L_2 regularization, and an Adam optimizer. The number of epochs for a training run was 100. Training was aimed at minimizing a multi-label cross entropy loss. The network was implemented using Keras with Tensorflow as the back-end.

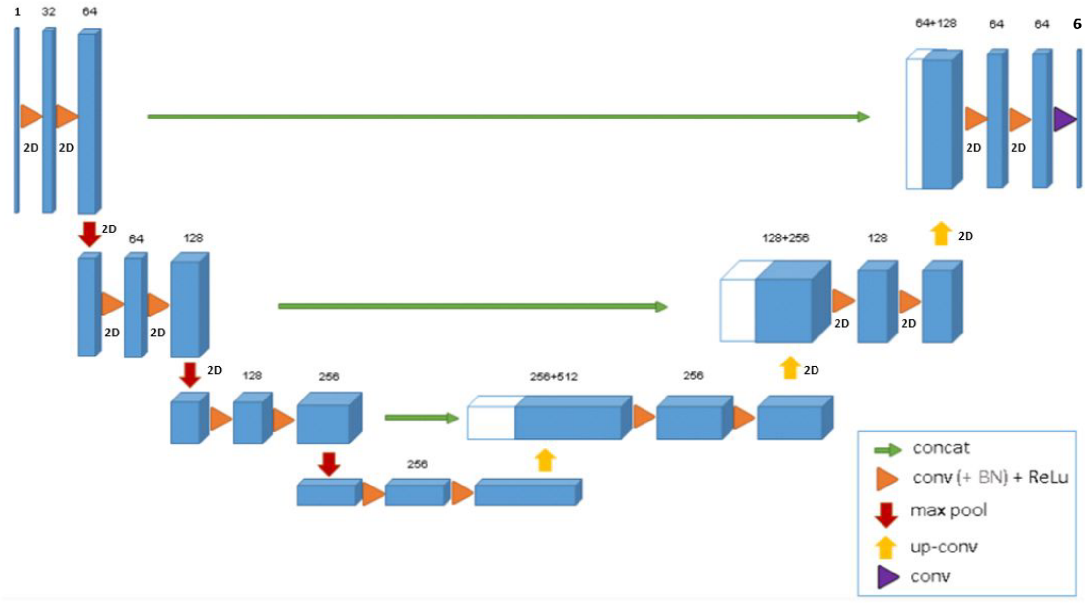


Figure 5.3: MRI-specific Deep Learning architecture used in this project. Architecture aniso-3DUNET developed by Mooij et al. [33] and originally based on Çiçek's [32] 3D U-net architecture, that considers the anisotropy of the input MRI volumes (stack of 2D images).

5.2.1 DICE score

DICE score measures the similarity between sets, i.e. X and Y. If the two sets are identical (i.e. they contain the same elements), the DICE score is equal to 1.0, while if X and Y have no elements in common, it is equal to 0.0.

$$dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{5.9}$$

This measurement is commonly used in image segmentation tasks for medical applications, in particular for comparing the algorithm output, in the test phase, against the reference segmentations, which are usually called ground-truths. Using Equation 5.9 for this purpose, X would be the volume of predicted segmentation probabilities and Y the volume of ground truth labels. The metric usually used to evaluate the performance of the deep learning model in both training and test phases is the average of the DICE scores of the entire training and test dataset respectively.

5.2.2 K-Fold Cross Validation

Cross-validation is a statistical method commonly used to estimate the performance of machine learning models. There are many versions of cross-validation, and the most commonly used is k-fold cross-validation. In k-fold cross-validation, the data set is randomly split into k parts, commonly addressed as folds, each containing an approximately equal number of data samples.

This process consists of k iterations. In the first iteration, the first fold is left out for validation, and the remaining k-1 folds form the training data. The first model is trained on that training data and then applied to the validation set to evaluate the performance of the model. This procedure is repeated k-times, alternating the fold used as test set in the iterations. This way all the data is used for training and for validation. However, the performance analysis remains honest since a training and the validation set separation is always guaranteed. An Illustration of this separation is shown in Figure 5.4.

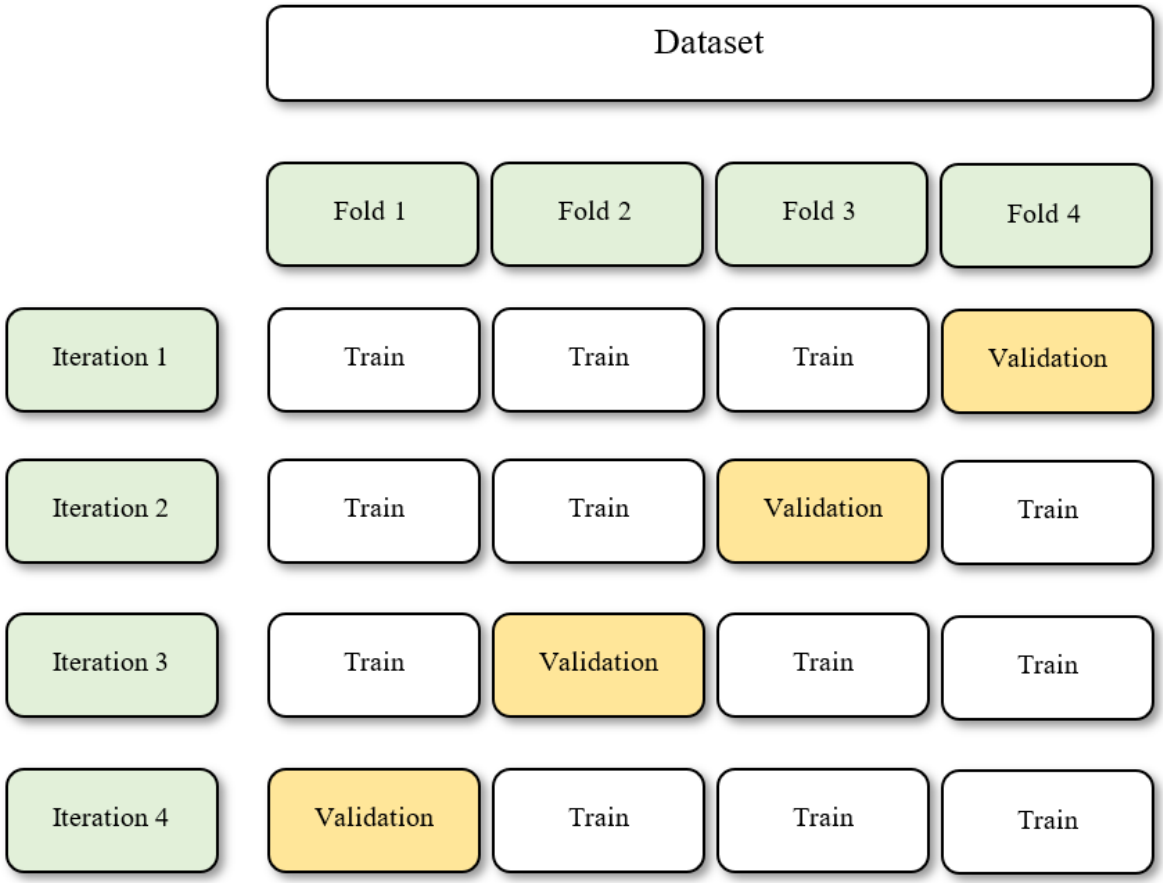


Figure 5.4: Overview of a k-fold cross validation example, where k = 4. The entire dataset is partioned into 4 groups of samples, called folds.

5.3 Dataset

Images from two databases were used in this study+, both are consecutive cohorts of diagnostic prostate-MRI, one is from an internal 2016-Archive and the other from the 2012 PROSTATEx challenge. The 2012-dataset was acquired on two different types of Siemens 3T MR-scanners, the MAGNETOM Trio and Skyra, using a turbo spin-echo sequence. The 2016-dataset was acquired on a Siemens 3T MR-scanner, the MAGNETOM Skyra using a turbo spin-echo sequence. All T2-weighted MR volumes used

had either a fixed voxel size of 0.5x0.5x3.6 mm or 0.3x0.3x3.6 mm. The T2-weighted images, from both databases, were acquired with several combinations of acquisition parameters (repetition time, echo time). The exclusion criteria applied were the absence of T2-weighted and DWI images in the database and the absence of any radiologist identified lesions in the patient's report.
Four different datasets were selected from the 2 databases mentioned.

1. Randomized subset of 32 T2-weighted MRI volumes from the 2016-dataset, manually annotated by the author of this dissertation using the segmentation tool ITK-SNAP [35], used to train the segmentation model and validate the multi-reference tissue normalization manual approach. The reference-tissues were segmented as shown in Figure 5.5.

2. Randomized subset of 8 T2-weighted MRI volumes from the 2016-dataset, manually annotated by the author of this dissertation, using the segmentation tool ITK-SNAP [35], used to test the segmentation model.

3. Randomized subset of 88 T2-weighted MRI volumes from the 2016-dataset used to validate the automation of our multi-reference tissue method.

4. Randomized subset of 49 T2-weighted MRI volumes from the 2012-dataset used to validate the automation of our multi-reference tissue method.

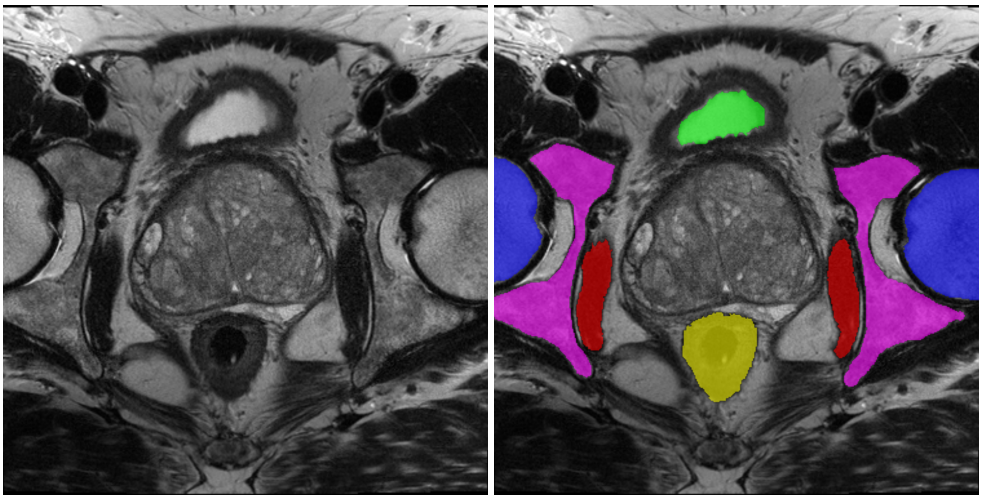


Figure 5.5: Original image (left) and manual segmentation overlap (right) of a T2-weighted MRI axial slice. The color representation is: Red - Obturator-Internus muscle; Green - Bladder lumen; Pink - Pelvic bone; Blue - Femur heads; Yellow - Rectum.

All patients' scans were analyzed by experienced radiologists using PI-RADS scoring to report the malignancy and localization of lesions found. These lesions were further studied and Gleason scores were obtained through biopsies. Gleason-6 and lower-scored lesions were classified as clinically non-significant (benign) prostate lesions, while Gleason-7 and higher-scored lesions were classified as clinically significant (malignant) prostate lesions. The Peripheral and Transition zone prostate lesions were delineated

by trained observers in all scans. These Gleason-scored lesions were delineated in T2-weighted scans with the help of Apparent diffusion coefficient maps (acquired with DWI) and the radiologist's report. To avoid any prevailing human errors in lesion segmentation and guaranty the correct region-of-interest, all lesions were 2-pixel eroded. The reference-tissue segmentation was performed by the author of this dissertation and approved by a radiologist. The reference-tissue segmentation task was time-expensive, taking 8 months to acquired all the segmentations used in this dissertation.

5.4 Experiments

To prove the dissertation initial hypothesis and additional assumptions made, several experiments were conducted:

5.4.1 Baseline - Manual Multi-Reference Tisssue Normalization

The method introduced by Stoilescu et al., in [7], applied PI-RADS scored lesions. However, in order to improve the reference standard, Biopsy Gleason scores, a more reliable measurement, were used in this dissertation. Subsequently, the results of the novel multi-reference tissue normalization manual method, using the first dataset (fully manual), were reported as baseline. The diagnostic accuracy was tested using ROC curve analysis before and after normalization. The two Areas Under the Curve (AUC) were compared.

A decrease in inter-sample intensity variability was expected after normalization. Thus, raw and normalized inter-patient variability was tested on all reference-tissues using an F-test for equality of variance, with a significance level of 0.05 ($\alpha = 0.05$). This possible decrease of intensity variability can cause the loss of sample-specific information. Therefore, the effect of the normalization on reference-tissue discriminability was also tested. Box-plots displaying the distribution of intensities, before and after normalization, in all reference-tissues were plotted to visually evaluate reference-tissue discriminability.

5.4.2 Linear Model vs Smooth Cubic Hermite Model

Throughout the development of this novel multi-reference tissue normalization method some assumptions were made, one of them being that a linear model, proposed in the past, could not fully predict the MR intensities non-linear distribution. To prove this assumption, a comparison between a piece-wise linear model and a piece-wise smooth cubic Hermite model was performed, using the first dataset (fully manual). The diagnostic accuracy of the two models was evaluated using Receiver Operating Characteristic. The two Areas Under the Curve were compared.

5.4.3 Validation of Deep Learning Model

The performance of the Deep Learning (DL) model was evaluated to guarantee it could be used to produce acceptable multi-tissue segmentations. The automatic output of the model and the manual segmentations, considered as the ground truth, were compared using DICE score analysis. Due to the low number of annotated volumes available for training in Dataset 1, the performance of the trained model was expected to unreliable when tested, so a 4-fold cross validation was carried out to obtain a range of training DICE scores that would be more representative of the predictive ability of the model. After the 4-fold cross-validation runs, one more model training was carried out on all the training volumes in Dataset 1 and predictions were generated on the test volumes of Dataset 2 that had been kept separate. A minimal DICE score equal or higher than 0.80 was considered, in a group discussion, to validate the trained DL model.

5.4.4 Automatic vs Manual Multi-Reference Tissue Normalization

The main hypothesis of this dissertation lies on the feasibility of automating the novel multi-reference tissue normalization method proposed. In order to prove this hypothesis, the manual (baseline) and automatic approaches were compared. The effect of the automatic normalization on diagnostic accuracy was first determined through ROC analysis, using the third and fourth datasets. The two Areas Under the Curve, before and after automatic normalization, were compared. Then, manual and automatic normalization were compared using the already performed manual and automatic ROC analyses and respective AUCs.

To further validate the automatic multi-reference tissue normalization method, the effect of the automatic normalization was studied on inter-sample variability and reference-tissue discriminability, using the combination of the third and fourth dataset. A decrease in inter-patient variability was expected and tested on all reference-tissues using an F-test for equality of variance, with a significance level of 0.05 ($\alpha = 0.05$). Box-plots displaying the distribution of intensities, before and after normalization, in all reference-tissues were plotted to visually evaluate reference-tissue discriminability.

5.4.5 Databases Normalization Effect

In the Dataset Section, the two databases used were introduced. Two distinct mp-MRI databases were implemented to test the effect of the normalization method in more than one database and solidify the findings of this study. The third (subset of 2016 database) and fourth datasets (subset of 2012 database) were expected to have different levels of machine variability, which would reflect on a different normalization effect in each dataset, with bigger variability reduction on datasets with more variability. To prove this assumption, diagnostic accuracy and inter-sample intensity variability were tested, in both datasets, and compared. The normalization effect on diagnostic accuracy was tested with an ROC analysis, before and after multi-reference tissue normalization, using AUC as the evaluation metric.

The inter-sample intensity variability was compared in the two datasets in their raw forms (before normalization), using an F-test for equality of variance, with a significance level of 0.05 ($\alpha = 0.05$).

5.4.6 Analysis of the Optimal Number of Reference-Tissues

The normalization method proposed in this dissertation was built on the basis that a linear model, proposed in the past, could not fully predict the MR intensities non-linear distribution. It was also assumed that the more information was provided the more accurate the normalization would be. Thus, using more reference tissues would allow better results.

To test this assumption, the reference-tissue normalization method was implemented using from 1 to 5 reference tissues, with all combinations of tissues used in this dissertation applied. In cases envolving only 1 or 2 reference tissues the interpolation method applied was linear interpolation as cubic Hermite interpolation can only be used when more data-points are input. ROC analyses, evaluating diagnostic accuracy after normalization, were performed for all combinations of 1 to 5 reference tissues. The AUCs of these analyses were all scatter plotted to illustrate how the number of reference-tissues applied can affect the reference-tissue normalization method.

5.4.7 Normalization Effect in PZ and TZ Lesions

T2-weighted MRI is considered the superior of all the mp-MRI sequences for detection of cancer in the Transition zone [15], while Peripheral zone lesions are usually classified using additional MRI sequences, such as Diffusion-Weighted MRI.

Two experiments were performed to evaluate TZ and PZ lesions individually. The first experiment studied the effect of automatic multi-reference tissue normalization on the differentiation of clinically significant TZ lesions and clinically non-significant TZ lesions. The second experiment studied the effect of automatic multi-reference tissue normalization on the differentiation of clinically significant PZ lesions and clinically non-significant PZ lesions. In both experiments, diagnostic accuracy was tested using ROC curve analysis before and after normalization. T2-weighted raw intensities were expected to have better diagnostic accuracy in TZ lesions compared to PZ lesions.

Chapter 6

Results and Discussion

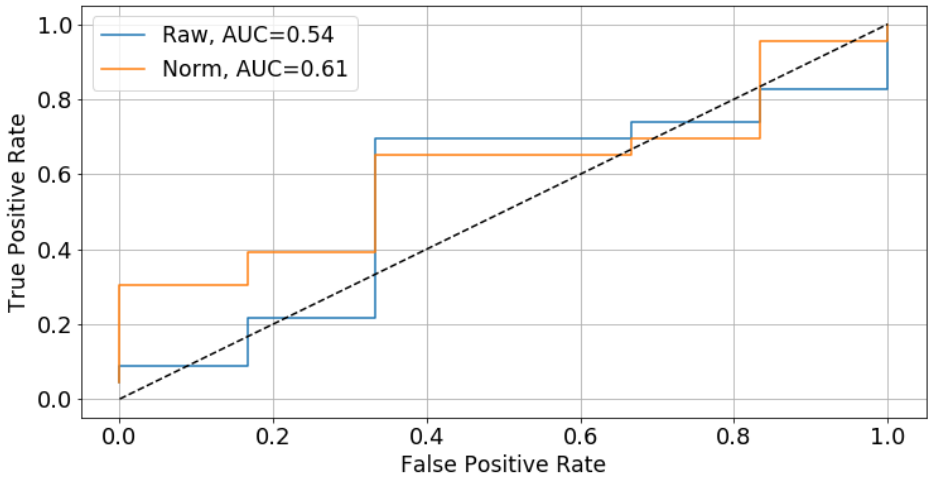
6.1 Baseline - Manual Multi-Reference Tisssue Normalization

The effect of the manual multi-reference tissue normalization method on diagnostic accuracy is displayed in Figure 6.1(b).

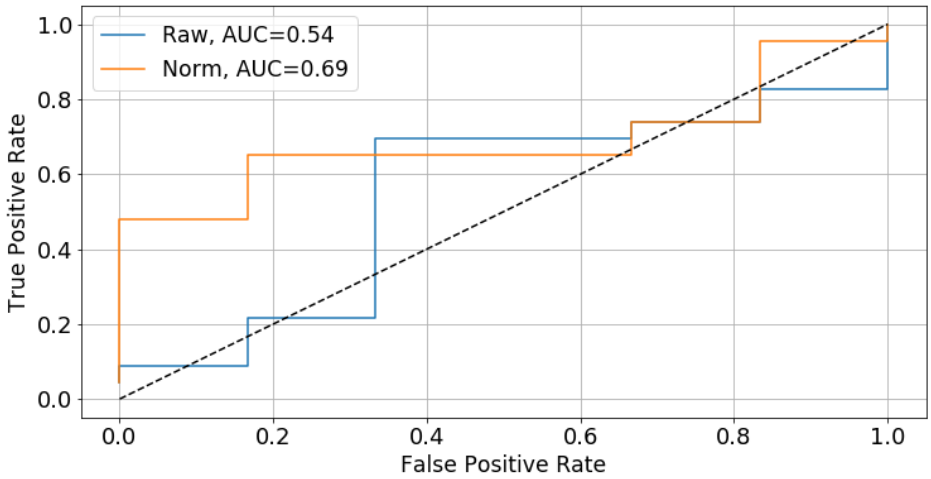
These results show a clear improvement in diagnostic accuracy from an AUC of 0.54 in raw T2-weighted intensities to an AUC of 0.69 in normalized T2-weighted intensities. This experiment confirmed that the manual multi-reference tissue normalization of Gleason scored lesions improved diagnostic accuracy, discriminating clinically non-significant lesions from clinically significant lesions. Stoilescu et al. in [5], report a diagnostic accuracy increase from $AUC_{RAW} = 0.66$ to $AUC_{NORM} = 0.89$ using a variant of this manual multi-reference tissue normalization method. These results are in line with the results shown in Figure 6.1(b), even with the difference in type of lesion score used. Therefore, the positive effect of this normalization method has also been validated for Gleason-scored lesions, which are more reliable than the PI-RADS scored lesions used in [5].

Inter-sample intensity variability was also tested before and after applying the manual multi-reference tissue normalization method. The results of the F-test applied, demonstrated a significant ($p_{value} < 0.05$) decrease in inter-patient variability on all reference tissues except the Obturator-Internus muscle. However, reducing inter-sample variability could mean also the loss of sample-specific information. Therefore, the inter-sample intensity distribution was plotted in Figure 6.2, to evaluate the effect of the normalization on reference-tissue discriminability.

From raw to normalized, the box-plots, displayed in Figure 6.2, showed more distinguishable reference-tissues and a clear decrease in inter-sample intensity variability. The box-plots in this Figure showed this



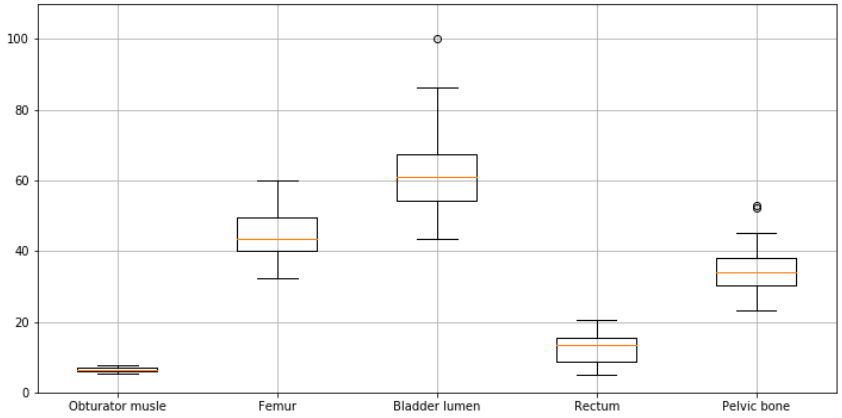
(a) Linear Interpolation



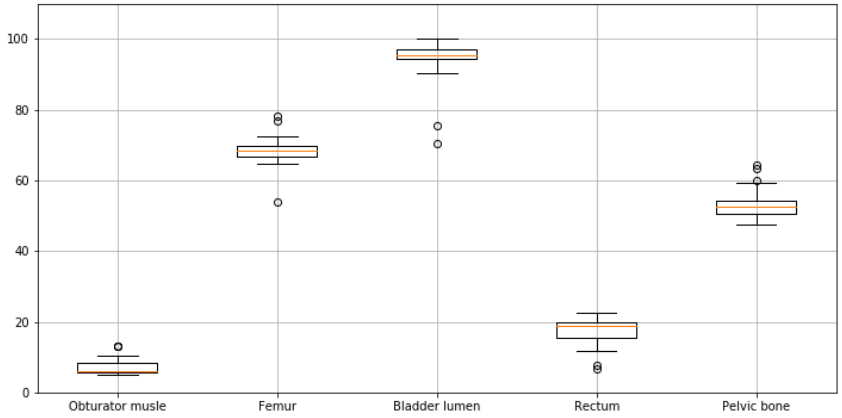
(b) Cubic Hermite Interpolation

Figure 6.1: ROC curves for differentiation of clinically significant and other prostate lesions before (blue) and after (orange) normalization, using (a) Linear and (b) Cubic Hermite interpolation

method does not only affect prostate intensities, but the entire intensity range covered by the reference tissues used, indicating that this method could possibly be generalized to other MR targeted body-areas. The results represented in Figure 6.2 also suggest it could finally be possible to distinguish tissues according to their intensity distribution, thus addressing a relevant MR-related issue, the lack of a tissue-specific value in MRI.



(a) Raw



(b) Normalized

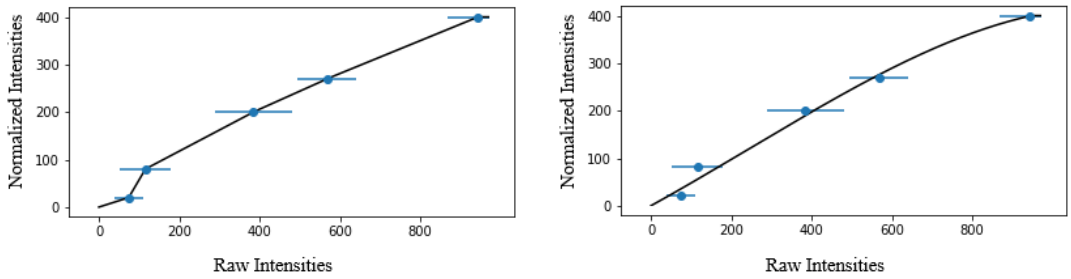
Figure 6.2: Box-plots showing the inter-sample distribution of intensities in each reference-tissue, before (a) and after (b) applying the manual multi-reference tissue normalization method, scaled to a range between 0-100.

6.2 Linear Model vs Smooth Cubic Hermite Model

Two types of interpolation methods were applied on the multi-reference tissue normalization method. In Figure 6.3, an example of the curve models acquired for the normalization of one of the T2-weighted scans is displayed for visualization purposes, using Linear and Cubic Hermite interpolation.

Figure 6.1 shows the ROC analyses on diagnostic accuracy of the Linear and Cubic Hermite interpolated models. These ROC analyses show an improvement on diagnostic ability using the non-linear (Cubic Hermite) approach, going from an AUC of 0.61 (Linear) to an AUC of 0.69 (Cubic Hermite).

These results can be explained by the fact that reference tissue measurements are noisy. Fitting exactly



(a) Linear Interpolation

(b) Cubic Hermite Interpolation

Figure 6.3: Curve models used for the normalization of one of the T2-weighted scans acquired through (a) Linear and (b) Cubic Hermite interpolation.

through the points provides a noisy non-linear curve, while a smooth fit using Cubic Hermite interpolation reduces noise by not fitting through the points.

These results can also prove the assumption that a linear model can not fully predict the MR intensities non-linear distribution, demonstrating that a continuously smooth normalization model better represents the data.

6.3 Validation of Deep Learning Model

The range of average DICE scores obtained with cross-validation and the average DICE scores acquired during the test phase are reported in Table 6.1.

Table 6.1: Range of average DICE scores obtained with cross-validation and the average DICE scores of each reference-tissue segmentation acquired during the test phase.

DICE score	O.Muscle	Bladder L.	Femur H.	Rectum	Pelvic B.
Cross-Val	0.83-0.89	0.92-0.97	0.93-0.96	0.85-0.9	0.89-0.9
Testing	0.88	0.93	0.96	0.84	0.92

The DICE scores at the end of all runs were mostly stable, as shown in the Appendix section, thus a short run of 100 epochs was sufficient to obtain these results. The results in Table 6.1 validate the DL output segmentations with all DICE score averages ≥ 0.80, in both cross-validation and test runs. This threshold was decided to be used as validation of this model in a group discussion, since an erosion step is implemented afterwards, which can eliminate most segmentation errors. For visualization purposes, an example of an automatic segmentation, output from the Deep Learning model, is shown in Figure 6.4. The Deep Learning model, as shown in Figure 6.4, due to the low number of annotated volumes available for training, can sometimes over-predict the segmentation of the reference tissues (e.g. Obturator-Internus muscle), meaning the limits of the reference tissues may not be well defined and the Deep

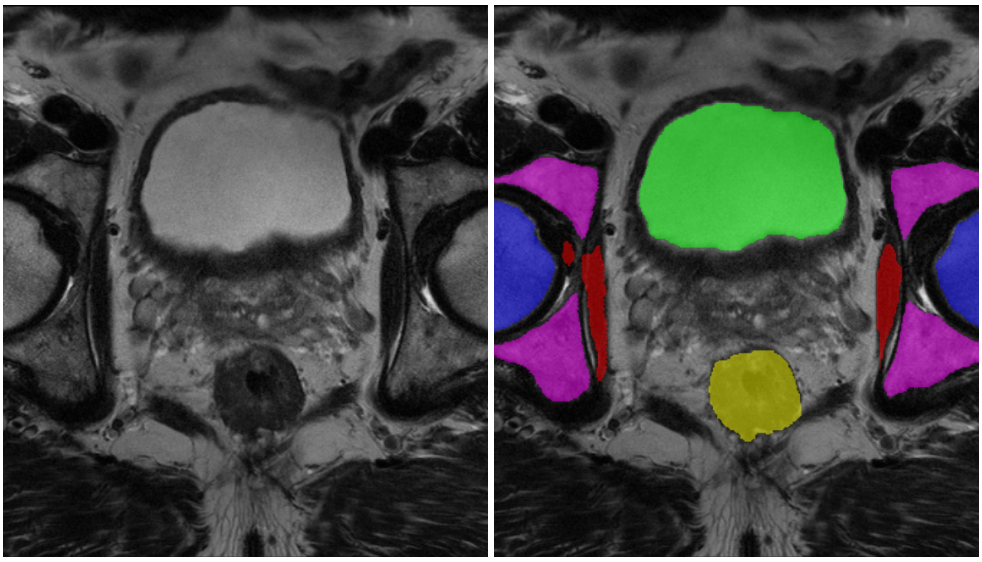


Figure 6.4: Original image (left) and automatic (DL output) segmentation overlap (right) of a T2-weighted MRI axial slice. The color representation is: Red - Obturator-Internus muscle; Green - Bladder lumen; Pink - Pelvic bone; Blue - Femur heads; Yellow - Rectum.

Learning Model segments over the boarders of these tissues. The erosion of reference tissues, applied before normalization, was introduced to account for this possible over-prediction and to guarantee the use of good regions of interest in the multi-reference tissue normalization method.

6.4 Auto vs Manual Multi-Reference Tissue Normalization

Figure 6.5 shows the results obtained using the automatic normalization approach. An improvement on diagnostic accuracy is visible when comparing raw and normalized images, going from an $AUC_{RAW} = 0.68$ to an $AUC_{NORM} = 0.73$.

Comparing manual and automatic ROC analyses, displayed in Figure 6.1(b) and Figure 6.5 respectively, some differences are noticeable. There is a larger normalization effect on diagnostic accuracy of T2-weighted images using the manual approach. However, a large difference of raw AUCs in the two datasets is also clear, with a smaller raw diagnostic accuracy in the T2-weighted images from Dataset 1. An initial raw diagnostic accuracy in T2-weighted images can reflect on the normalization effect. For example, a poor initial raw diagnostic accuracy can lead to bigger increases of performance, while a higher initial raw diagnostic accuracy could mean that there is less need for normalization and its effect will be less noticeable. This assumption can explain the difference in manual and automatic results.

Inter-sample intensity variability was also tested before and after applying the automatic multi-reference tissue normalization method. The results of the F-test applied, like the manual results in Experiment 6.1, demonstrated a significant ($p_{value} < 0.05$) decrease in inter-patient variability on all reference tissues

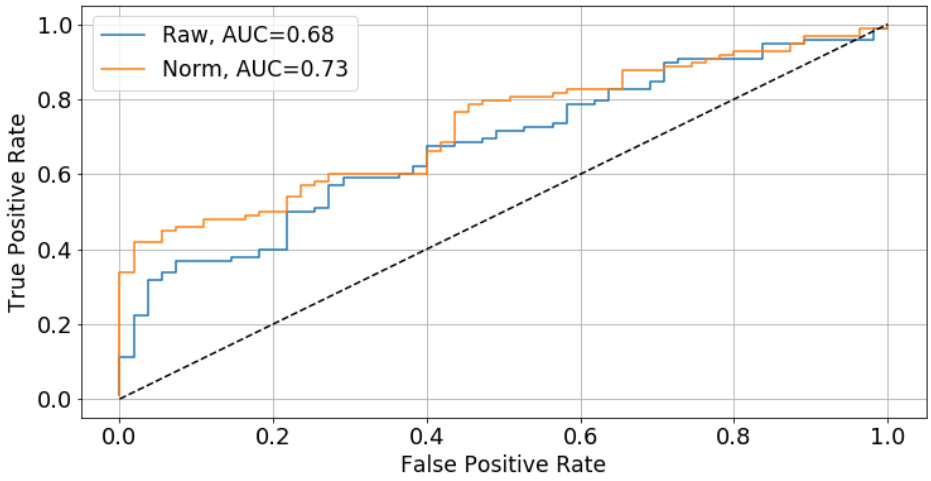


Figure 6.5: ROC curves for differentiation of clinically significant and other prostate lesions before (blue) and after (orange) normalization, using the automatic normalization approach

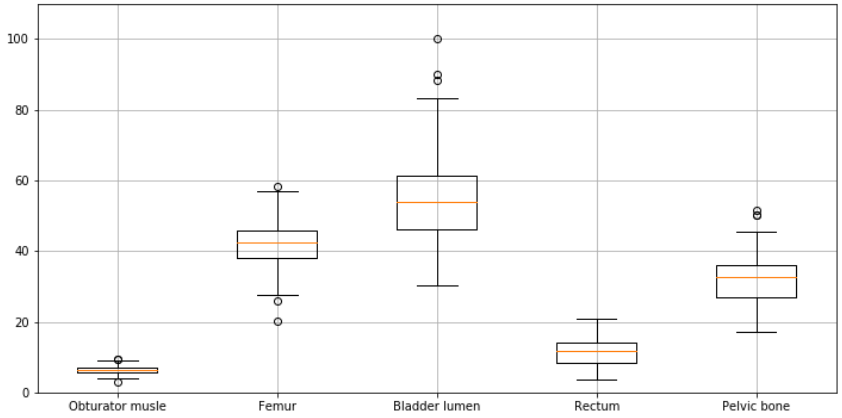
except the Obturator-Internus muscle. However, as mentioned before, the reduction of inter-sample variability could mean also the loss of sample-specific information. Therefore, the inter-sample intensity distribution was plotted in Figure 6.2, to evaluate the effect of the normalization on reference-tissue discriminability.

From raw to normalized, the box-plots, displayed in Figure 6.6, showed, like in the manual experiment, a clear decrease in inter-sample intensity variability (already proven above in experiment 6.1) and more distinguishable reference-tissues. However, the normalized reference tissues are more distinguishable applying the manual approach (Figure 6.6(b)) compared to the automatic approach (Figure 6.2(b)).

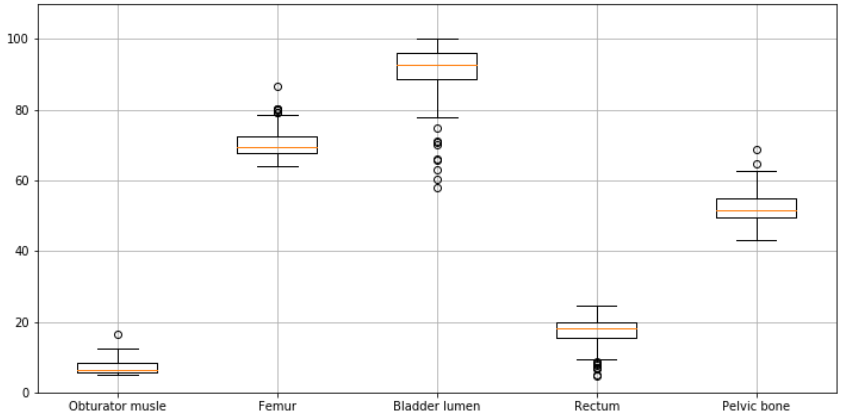
The automatic multi-reference tissue normalization method was shown to improve quantification of T2-weighted images (Figure 6.6) and diagnostic accuracy (Figure 6.5), possibly leading to a decrease in radiologist's interpretation variability. This research is in line with Peng's [6] results and conclusions, both confirming that T2-weighted image intensities are correlated to lesion malignancy [36] and that reference-tissue normalization methods have a positive impact on prostate cancer diagnostic accuracy [27].

6.5 Databases Normalization Effect

Figure 6.7 shows that the multi-reference tissue normalization method has a larger effect in the 2012-dataset (fourth dataset). Peng [6] who used the one-reference tissue normalization method also reported a different normalization effect on two databases. One of the raw T2-weighted databases reported bet-



(a) Raw



(b) Normalized

Figure 6.6: Box-plots showing the inter-sample distribution of intensities in each reference-tissue, before (a) and after (b) applying the automatic multi-reference tissue normalization method, scaled to a range between 0-100.

ter diagnostic ability compared to the other. However, only the database with the lower initial (raw) diagnostic accuracy had an improvement after normalization.

In this dissertation, the considerable different effect of reference-tissue normalization between the two datasets could be attributed to the fact that the 2012-dataset was acquired using two MR-scanners and was older than the 2016-dataset, possibly having more intensity variability. However, the intensity variability analysis with an F-test showed no significant ($p_{value} > 0.05$) differences between the two datasets.

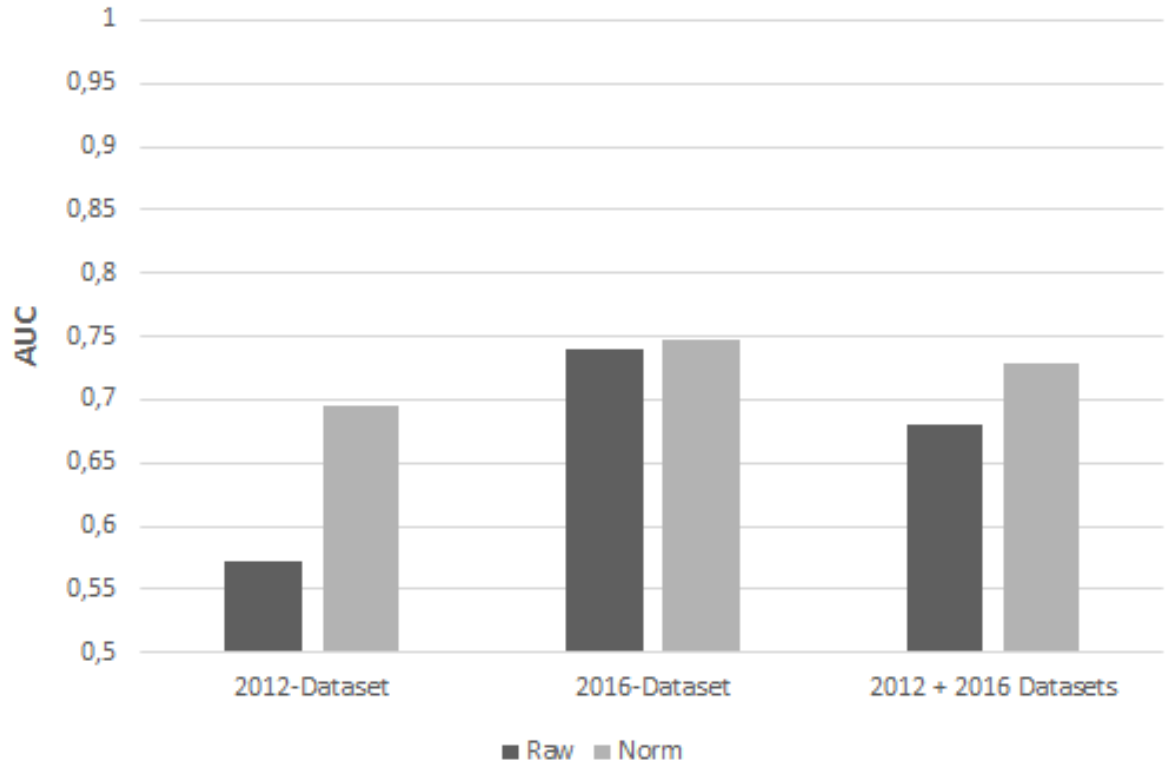


Figure 6.7: Histogram displaying the effect of multi-reference tissue normalization on diagnostic accuracy in the different datasets (Dataset 3 and 4). Considering that an AUC of 0.5 represents a random classifier, the results are displayed with AUC = 0.5 as a starting point.

6.6 Analysis of the Optimal Number of Reference-Tissues

Diagnostic accuracy of normalization approaches applying all combinations of 1 to 5 reference tissues were tested using ROC analyses. The AUCs of these analyses were all scatter plotted, in Figure 6.8, to illustrate how the number of reference-tissues applied can affect the reference-tissue normalization method.

In Figure 6.8, an increase of the number of reference tissues implemented in the normalization method lead to a small but noticeable increase of diagnostic accuracy, measured using AUC. This small increase in diagnostic accuracy does not constitute strong evidence that the higher the number of reference tissues the better the diagnostic performance acquired after normalization. However, it can be already an indicative that this is the case.

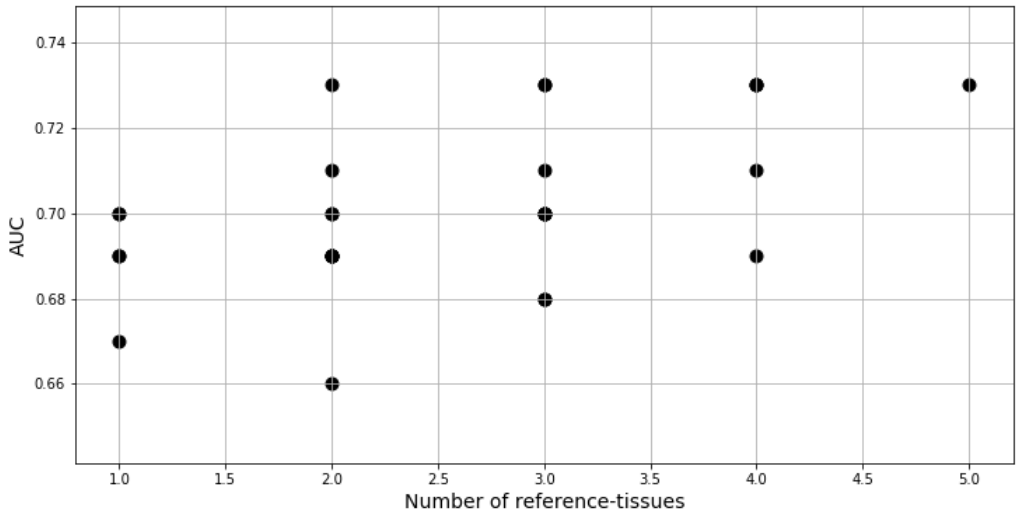


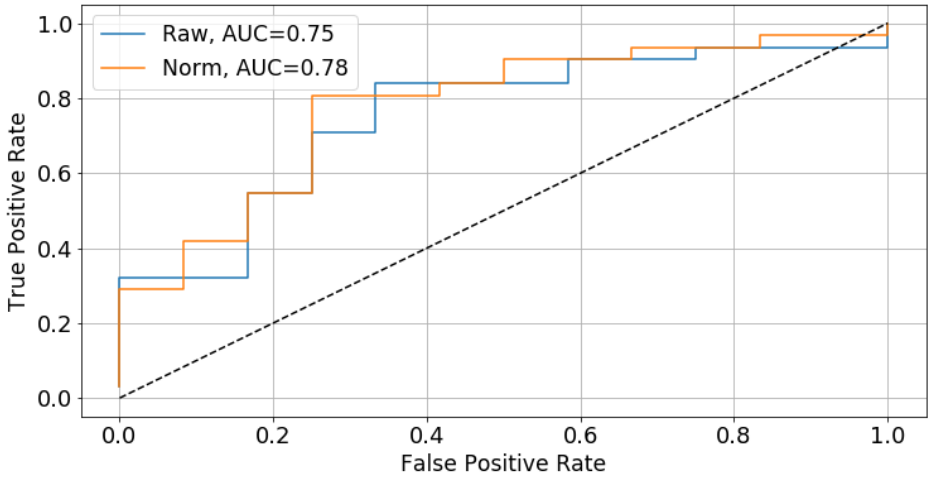
Figure 6.8: Diagnostic accuracy, measured by the AUC, scatter plotted against the number of reference-tissues implemented in the normalization method.

6.7 Normalization Effect in PZ and TZ Lesions

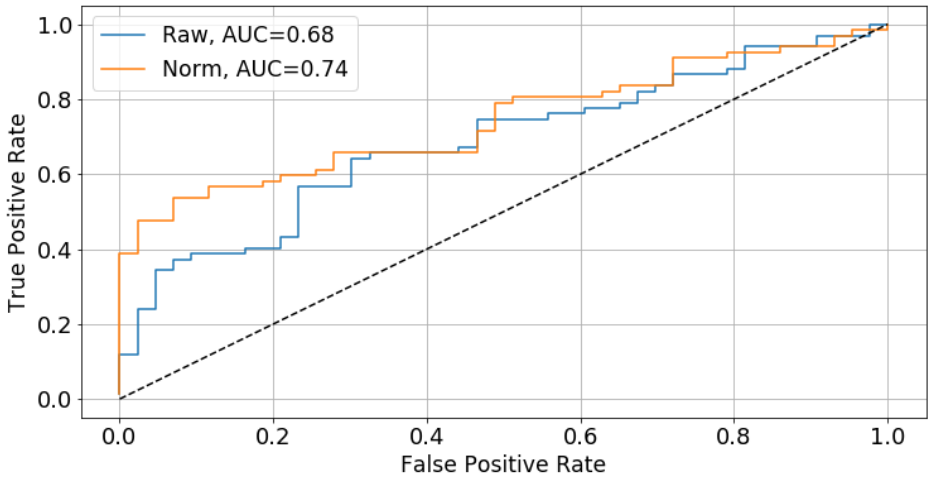
The results of the two experiments performed to individually evaluate the normalization effect on Peripheral and Transition zone lesions are shown in Figure 6.9.

The ROC analyses show, as expected, a higher diagnostic accuracy of raw T2-weighted intensities in TZ lesions compared to PZ lesions. These results are explained by the fact that T2-weighted MRI is currently considered the superior of all the mp-MRI sequences for detection of cancer in the Transition zone [15], while Peripheral zone lesion classification usually requires additional MRI sequences, such as Diffusion-Weighted MRI.

After multi-reference tissue normalization the diagnosis of both lesion types (PZ and TZ) improves. There is a larger normalization effect on PZ lesion diagnostic accuracy of T2-weighted images. However, a large difference between raw AUCs is also clear, with a smaller raw PZ lesion diagnostic accuracy. As mentioned above, the initial raw diagnostic accuracy in T2-weighted images can reflect on the normalization effect. For example, a poor initial raw diagnostic accuracy can lead to larger increases of performance, while a higher initial raw diagnostic accuracy could mean that there is less need for normalization and its effect will be less noticeable. The PZ lesion classification accuracy increased to TZ lesion classification level, which could potentially mean that the use of additional MR sequences would no longer be required for PZ lesion diagnosis. However, this bold assumption can only be supported with further research.



(a) Transition Zone Lesions



(b) Peripheral Zone Lesions

Figure 6.9: ROC curves for differentiation of clinically significant and other prostate lesions, (a) Transition zone lesions, (b) Peripheral zone lesions, before (blue) and after (orange) normalization.

Chapter 7

Conclusions

In this study, it was first confirmed (Figure 6.3) that the manual multi-reference tissue normalization of Gleason scored lesions improved diagnostic accuracy, discriminating Gleason-6 and lower lesions from Gleason-7 and higher lesions, with bigger improvements using a smooth interpolant (Cubic Hermite) compared to a linear interpolant, as expected.

Then, the application of a 3D variant of U-net [32] on the segmentation of multiple Pelvic tissues in 3D MRI images was reported and validated (Table 6.1). The automatic multi-reference tissue normalization approach, like the manual, was tested and showed to reduce inter-patient intensity variability and improve the separation of reference-tissues in intensity scale. Finally, the main hypothesis of this dissertation was corroborated with the latter results and Figure 6.5, which proved that it is possible to reproduce the manual results using the automatic normalization approach.

These results corroborate main hypothesis of this dissertation which consists on the feasibility to automate a novel multi-reference tissue normalization method.

It is possible to conclude that the developed novel normalization method (both manual and automatic approaches) improves quantification of T2-weighted images and diagnostic accuracy, possibly leading to a decrease in radiologist's interpretation variability. Moreover, the results presented in this dissertation provide strong evidence that T2-weighted image intensities are correlated to lesion malignancy [36] and that reference-tissue normalization methods have a positive impact on prostate cancer diagnostic accuracy [6].

Finally, the author of this dissertation concludes the novel automatic multi-reference normalization method can be applied in a clinical setting using the automatic approach, however more improvements in technical details need to be applied. Some of these details are explained below.

Prostate cancer is one of the current major societal challenges in health care and the application of this multi-reference normalization method can be a crucial factor in achieving precise determination of prostate lesion malignancy, which in turn could lead to correct treatment choices, preventing over-diagnosis and over-treatment. However, to reach this goal, the author of this dissertation still suggests the need for further research and improvements.

To try to mitigate the shortcomings of this dissertation an improvement of this work would consist of applying the normalization method in scans acquired from multiple vendors, instead of only a single MR manufacturer, to provide stronger evidence of this method's effectiveness.

Another improvement of this work would be to apply a larger dataset for training of the segmentation model which would lead to an improvement of the predictive performance of the trained Deep Learning model thus discarding the need for an erosion method before normalization.

Furthermore, since it has been proven in the past that segmentation methods benefit from intensity normalization [24, 37], an interesting future experiment, would be to input the normalized images in the deep learning trained model to investigate possible improvements in its predictive accuracy.

Finally, it would be interesting to investigate the potential of the multi-reference tissue normalization method on other MR-imaged parts of the body, since all this methodology requires is an appropriate set of not-commonly cancerous, homogeneous reference-tissues.

References

[1] R. L. Siegel, K. D. Miller, and J. Ahmedin, "Cancer statistics, 2018," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30, 2018.

[2] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley, M. J. Roobol, S. Eggener, M. Ghei, A. Villers, F. Bladou, G. M. Villeirs, J. Virdi, S. Boxler, G. Robert, P. B. Singh, W. Venderink, B. A. Hadaschik, A. Ruffion, J. C. Hu, D. Margolis, S. Crouzet, L. Klotz, S. S. Taneja, P. Pinto, I. Gill, C. Allen, F. Giganti, A. Freeman, S. Morris, S. Punwani, N. R. Williams, C. Brew-Graves, J. Deeks, Y. Takwoingi, M. Emberton, and C. M. Moore, "MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis," *The New England Journal of Medicine*, vol. 378, no. 19, pp. 1767–1777, 2018.

[3] A. B. Rosenkrantz, L. A. Ginocchio, D. Cornfeld, A. T. Froemming, R. T. Gupta, B. Turkbey, A. C. Westphalen, J. S. Babb, and D. J. Margolis, "Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists," *Radiology*, vol. 280, no. 3, pp. 793–804, 2016. PMID: 27035179.

[4] A. B. Rosenkrantz, R. P. Lim, M. Haghighi, M. B. Somberg, J. S. Babb, and S. S. Taneja, "Comparison of Interreader Reproducibility of the Prostate Imaging Reporting and Data System and Likert Scales for Evaluation of Multiparametric Prostate MRI," *American Journal of Roentgenology*, vol. 201, no. 4, pp. W612–W618, 2013.

[5] L. Stoilescu and H. Huisman, "Feasibility of multireferencetissue normalization of T2-weighted prostate MRI," in *Radiological Society of North America 2017 Scientific Assembly and Annual Meeting*, (McCormick Place Convention Center, Chicago), 2017.

[6] Y. Peng, Y. Jiang, and A. Oto, "Reference-tissue correction of T_2-weighted signal intensity for prostate cancer detection," in *Medical Imaging 2014: Computer-Aided Diagnosis*, vol. 9035, p. 903508, 2014.

[7] L. Stoilescu, M. C. Maas, and H. Huisman, "Feasibility of multireference tissue normalization of T2-weighted prostate MRI," in *European Society for Magnetic Resonance in Medicine and Biology*, (Palau de Congressos, Barcelona), 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 195, pp. 234–241, Springer International Publishing, 2015.

[9] L. Maaske, "Illustrated male reproductive system anatomy," in *http://medimagery.com/illustrated-male-reproductive-anatomy/*, 2010.

[10] J. E. McNeal, "The zonal anatomy of the prostate," *The Prostate*, vol. 2, no. 1, pp. 35–49, 1981.

[11] Y. J. Choi, J. K. Kim, N. Kim, K. W. Kim, E. K. Choi, and K.-S. Cho, "Functional MR Imaging of Prostate Cancer," *RadioGraphics*, vol. 27, no. 1, pp. 63–75, 2007. PMID: 17234999.

[12] P. A. Pollock, A. Ludgate, and R. J. Wassersug, "In 2124, half of all men can count on developing prostate cancer," *Current Oncology*, vol. 22, no. 1, p. 10–12, 2015.

[13] L.-Y. GL, A. PC, M. DF, and et al, "Outcomes of localized prostate cancer following conservative management," *Journal of the American Medical Association*, vol. 302, no. 11, pp. 1202–1209, 2009.

[14] C. M. Hoeks, M. G. Schouten, J. G. Bomers, S. P. Hoogendoorn, C. A. H. van de Kaa, T. Hambrock, H. Vergunst, J. M. Sedelaar, J. J. Fütterer, and J. O. Barentsz, "Three-tesla magnetic resonance–guided prostate biopsy in men with increased prostate-specific antigen and repeated, negative, random, systematic, transrectal ultrasound biopsies: Detection of clinically significant prostate cancers," *European Urology*, vol. 62, no. 5, pp. 902 – 909, 2012.

[15] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany, H. C. Thoeny, and S. Verma, "PI-RADS Prostate Imaging – Reporting and Data System: 2015, Version 2," *European Urology*, vol. 69, no. 1, p. 16–40, 2016.

[16] J. J. Fütterer, S. W. T. P. J. Heijmink, T. W. J. Scheenen, J. Veltman, H. J. Huisman, P. Vos, C. A. H. de Kaa, J. A. Witjes, P. F. M. Krabbe, A. Heerschap, and J. O. Barentsz, "Prostate Cancer Localization with Dynamic Contrast-enhanced MR Imaging and Proton MR Spectroscopic Imaging," *Radiology*, vol. 241, no. 2, pp. 449–458, 2006. PMID: 16966484.

[17] N. B. Delongchamps, M. Rouanne, T. Flam, F. Beuvon, M. Liberatore, M. Zerbib, and F. Cornud, "Multiparametric magnetic resonance imaging for the detection and localization of prostate cancer: combination of T2-weighted, dynamic contrast-enhanced and diffusion-weighted imaging," *British Journal of Urology International*, vol. 107, no. 9, pp. 1411–1418, 2010.

[18] J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager, and J. J. Fütterer, "ESUR prostate MR guidelines 2012," *European Radiology*, vol. 22, no. 4, pp. 746–757, 2012.

[19] J. Thompson, P. van Leeuwen, D. Moses, R. Shnier, P. Brenner, W. Delprado, M. Pulbrook, M. Böhm, A. Haynes, A. Hayen, and P. Stricker, "The Diagnostic Performance of Multiparametric Magnetic Resonance Imaging to Detect Significant Prostate Cancer," *The Journal of Urology*, vol. 195, no. 5, pp. 1428 – 1435, 2016.

[20] M. Kasel-Seibert, T. Lehmann, R. Aschenbach, F. V. Guettler, M. Abubrig, M.-O. Grimm, U. Teichgraeber, and T. Franiel, "Assessment of PI-RADS v2 for the Detection of Prostate Cancer," *European Journal of Radiology*, vol. 85, no. 4, pp. 726 – 731, 2016.

[21] R. Chou, J. Croswell, T. Dana, C. Bougatsos, I. Blazina, R. Fu, K. Gleitsmann, H. Koenig, C. Lam, A. Maltz, J. Rugge, and K. Lin, "Screening for prostate cancer: A review of the evidence for the U.S. Preventive Services Task Force," *Annals of Internal Medicine*, vol. 155, no. 11, pp. 762–771, 2011.

[22] Y. Zhuge, J. K. Udupa, J. Liu, and P. K. Saha, "Image background inhomogeneity correction in mri via intensity standardization," *Computerized Medical Imaging and Graphics*, vol. 33, no. 1, pp. 7 – 16, 2009.

[23] Y. Ge, J. K. Udupa, L. G. Nyúl, L. Wei, and R. I. Grossman, "Numerical tissue characterization in ms via standardization of the mr image intensity scale," *Journal of Magnetic Resonance Imaging*, vol. 12, no. 5, pp. 715–721, 2000.

[24] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[25] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Evaluating intensity normalization on MRIs of human brain with multiple sclerosis," *Medical Image Analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[26] N. Robitaille, A. Mouiha, B. Crépeault, F. Valdivia, S. Duchesne, and T. A. D. N. Initiative, "Tissue-Based MRI Intensity Standardization: Application to Multicentric Datasets," *International Journal of Biomedical Imaging*, vol. 2012, no. 347120, pp. 4–15, 2012.

[27] P. C. Vos, T. Hambrock, J. O. Barenstz, and H. J. Huisman, "Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI," *Physics in Medicine & Biology*, vol. 55, no. 6, pp. 1719–1734, 2010.

[28] K. K. Leung, M. J. Clarkson, J. W. Bartlett, S. Clegg, C. R. Jack, M. W. Weiner, N. C. Fox, and S. Ourselin, "Robust atrophy rate measurement in alzheimer's disease using multi-site serial mri: Tissue-specific intensity normalization and parameter selection," *NeuroImage*, vol. 50, no. 2, pp. 516 – 523, 2010.

[29] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[30] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 2146–2153, 2009.

[31] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*. O'Reilly Media, Inc.,1005 Gravenstein Highway North, Sebastopol, CA 95472, 2017.

[32] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* (S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, eds.), (Cham), pp. 424–432, Springer International Publishing, 2016.

[33] G. Mooij, I. Bagulho, and H. Huisman, "Automatic segmentation of prostate zones," *CoRR*, vol. abs/1806.07146, 2018.

[34] K. Padgett, A. Pollack, R. Stoyanova, A. Swallen, and A. Nelson, "Su-f-j-171: Robust atlas based segmentation of the prostate and peripheral zone regions on mri utilizing multiple mri system vendors," *Medical Physics*, vol. 43, pp. 3447–3447, 6 2016.

[35] P. A. Yushkevich, J. Piven, H. Cody Hazlett, R. Gimpel Smith, S. Ho, J. C. Gee, and G. Gerig, "User-Guided 3D Active Contour Segmentation of Anatomical Structures: Significantly Improved Efficiency and Reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.

[36] A. H. Dinh, R. Souchon, C. Melodelima, F. Bratan, F. Mège-Lechevallier, M. Colombel, and O. Rouvière, "Characterization of prostate cancer using T2 mapping at 3T: A multi-scanner study," *Diagnostic and Interventional Imaging*, vol. 96, no. 4, pp. 365 – 372, 2015.

[37] S. Roy, A. Carass, and J. L. Prince, "Patch based intensity normalization of brain MR images," in *2013 IEEE 10th International Symposium on Biomedical Imaging*, vol. 2013, pp. 342–345, April 2013.

Appendix

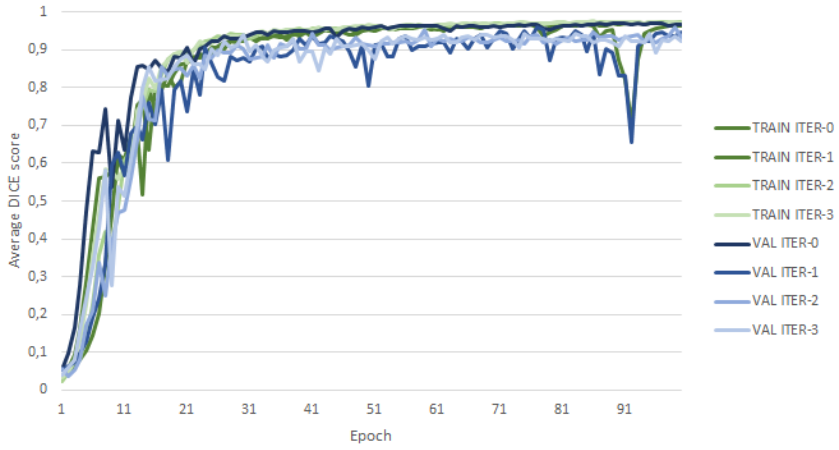


Figure 7.1: Dice scores achieved throughout the training and 4-fold cross-validation runs for the Bladder lumen segmentation. Green curves are training (TRAIN) and blue curves are validation (VAL) scores, and the colour tone is varied for each of the 4 cross-validation iteration.

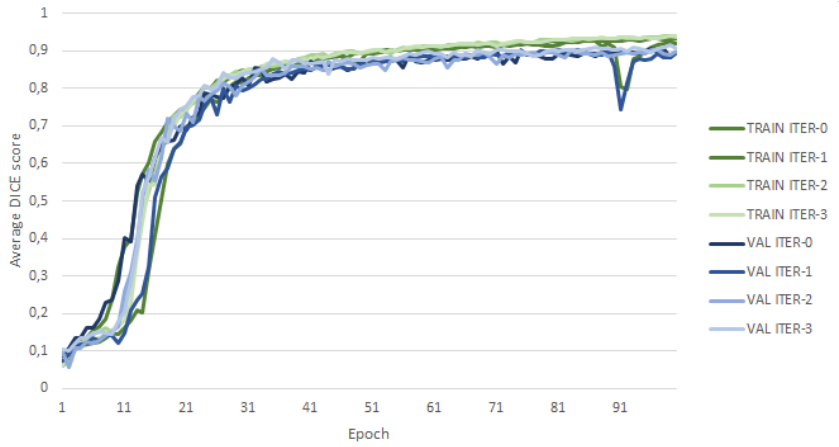


Figure 7.2: Dice scores achieved throughout the training and 4-fold cross-validation runs for the Pelvic bone segmentation. Green curves are training (TRAIN) and blue curves are validation (VAL) scores, and the colour tone is varied for each of the 4 cross-validation iteration.

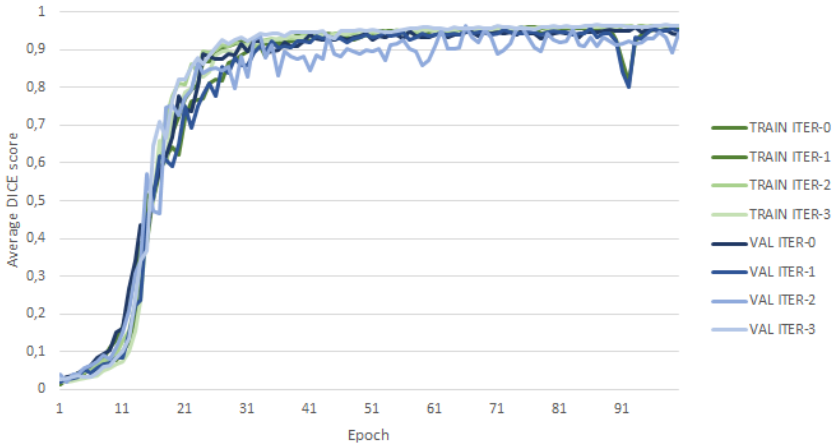


Figure 7.3: Dice scores achieved throughout the training and 4-fold cross-validation runs for the Femur heads segmentation. Green curves are training (TRAIN) and blue curves are validation (VAL) scores, and the colour tone is varied for each of the 4 cross-validation iteration.

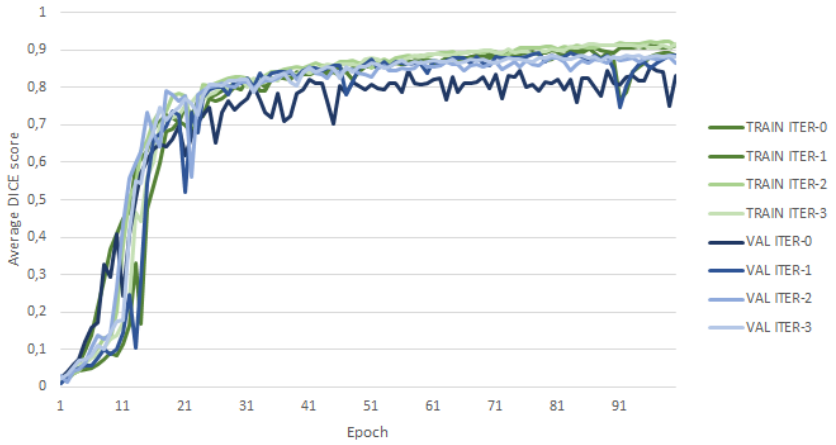


Figure 7.4: Dice scores achieved throughout the training and 4-fold cross-validation runs for the Obturator-Internus muscle segmentation. Green curves are training (TRAIN) and blue curves are validation (VAL) scores, and the colour tone is varied for each of the 4 cross-validation iteration.

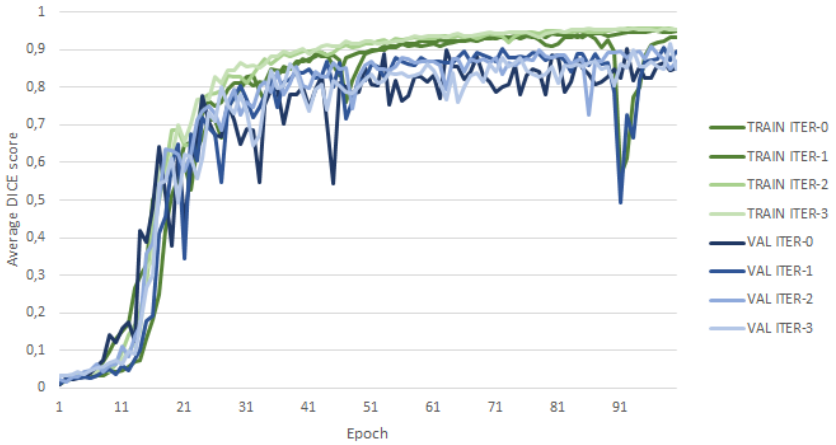


Figure 7.5: Dice scores achieved throughout the training and 4-fold cross-validation runs for the Rectum segmentation. Green curves are training (TRAIN) and blue curves are validation (VAL) scores, and the colour tone is varied for each of the 4 cross-validation iteration.